MEASURING STATISTICAL DEPENDENCE AND ITS APPLICATIONS IN MACHINE LEARNING

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Ze Jin

August 2018

MEASURING STATISTICAL DEPENDENCE AND ITS APPLICATIONS IN MACHINE LEARNING

Ze Jin, Ph.D.

Cornell University 2018

My PhD research focuses on measuring and testing mutual dependence and conditional mean dependence, and applying it to Machine Learning problems, which is elaborated in the following four chapters:

Chapter 1 – We propose three new measures of mutual dependence between multiple random vectors. Each measure is zero if and only if the random vectors are mutually independent. The first generalizes distance covariance from pairwise dependence to mutual dependence, while the other two measures are sums of squared distance covariances. The proposed measures share similar properties and asymptotic distributions with distance covariance, and capture non-linear and non-monotone mutual dependence between the random vectors. Inspired by complete and incomplete V-statistics, we define empirical and simplified empirical measures as a trade-off between the complexity and statistical power when testing mutual independence. The implementation of corresponding tests is demonstrated by both simulation results and real data examples.

Chapter 2 – We apply both distance-based and kernel-based mutual dependence measures to independent component analysis (ICA), and generalize dCovICA to MDMICA, minimizing empirical dependence measures as an objective function in both deflation and parallel manners. Solving this minimization problem, we introduce Latin hypercube sampling (LHS), and a global optimiza-

tion method, Bayesian optimization (BO) to improve the initialization of the Newton-type local optimization method. The performance of MDMICA is evaluated in various simulation studies and an image data example. When the ICA model is correct, MDMICA achieves competitive results compared to existing approaches. When the ICA model is misspecified, the estimated independent components are less mutually dependent than the observed components using MDMICA, while they are prone to be even more mutually dependent than the observed components using other approaches.

Chapter 3 – Independent component analysis (ICA) decomposes multivariate data into mutually independent components (ICs). The ICA model is subject to a constraint that at most one of these components is Gaussian, which is required for model identifiability. Linear non-Gaussian component analysis (LNGCA) generalizes the ICA model to a linear latent factor model with any number of both non-Gaussian components (signals) and Gaussian components (noise), where observations are linear combinations of independent components. Although the individual Gaussian components are not identifiable, the Gaussian subspace is identifiable. We introduce an estimator along with its optimization approach in which non-Gaussian and Gaussian components are estimated simultaneously, maximizing the discrepancy of each non-Gaussian component from Gaussianity while minimizing the discrepancy of each Gaussian component from Gaussianity. When the number of non-Gaussian components is unknown, we develop a statistical test to determine it based on resampling and the discrepancy of estimated components. Through a variety of simulation studies, we demonstrate the improvements of our estimator over competing estimators, and we illustrate the effectiveness of our test to determine the number of non-Gaussian components. Further, we apply our method to real data exam-

ples and show its practical value.

Chapter 4 – A crucial problem in statistics is to decide whether additional variables are needed in a regression model. We propose a new multivariate test to investigate the conditional mean independence of Y given X conditioning on some known effect Z, i.e., $\mathrm{E}(Y|X,Z) = \mathrm{E}(Y|Z)$. Assuming that $\mathrm{E}(Y|Z)$ and Z are linearly related, we reformulate an equivalent notion of conditional mean independence through transformation, which is approximated in practice. We apply the martingale difference divergence (MDD) to measure conditional mean dependence, and show that the estimation error from approximation is negligible, as it has no impact on the asymptotic distribution of the test statistic under some regularity assumptions. The implementation of our test is demonstrated by both simulations and a financial data example.

BIOGRAPHICAL SKETCH

Ze Jin was born in Hangzhou, Zhejiang, China in 1989.

Ph.D. Statistics, Cornell University 2018.

M.S. Statistics, Cornell University, 2016.

M.S. Statistics, Stanford University, 2013.

B.S. Statistics, Zhejiang University, 2011.

To my parents:

Xuefeng Jin and Shuping Yu

ACKNOWLEDGEMENTS

I want to thank my advisor, my committee, my professors, my colleagues, my families, and my friends, who have been helping me, encouraging me, supporting me, and with me for the last five years.

TABLE OF CONTENTS

CHAPTER 1

GENERALIZING DISTANCE COVARIANCE TO MEASURE AND TEST MULTIVARIATE MUTUAL DEPENDENCE VIA COMPLETE AND INCOMPLETE V-STATISTICS

Generalizing Distance Covariance to Measure and Test Multivariate Mutual Dependence via Complete and Incomplete V-Statistics

Ze Jin, David S. Matteson[1]

Abstract

We propose three new measures of mutual dependence between multiple random vectors. Each measure is zero if and only if the random vectors are mutually independent. The first generalizes distance covariance from pairwise dependence to mutual dependence, while the other two measures are sums of squared distance covariances. The proposed measures share similar properties and asymptotic distributions with distance covariance, and capture non-linear and non-monotone mutual dependence between the random vectors. Inspired by complete and incomplete V-statistics, we define empirical and simplified empirical measures as a trade-off between the complexity and statistical power when testing mutual independence. The implementation of corresponding tests is demonstrated by both simulation results and real data examples.

Key words: characteristic functions; distance covariance; multivariate analysis; mutual independence; V-statistics

1. Introduction

Let $X = (X_1, \ldots, X_d)$ be a set of variables where each component X_j, $j = 1, \ldots, d$ is a random vector, and let $\mathbf{X} = \{X^k = (X_1^k, \ldots, X_d^k) : k = 1, \ldots, n\}$ be an i.i.d. sample from F_X, the joint distribution of X. We are interested in testing the hypothesis

$$H_0 : X_1, \ldots, X_d \text{ are mutually independent}, \quad H_A : X_1, \ldots, X_d \text{ are dependent},$$

which has many applications, including independent component analysis [16, 26], graphical models [8, 10, 22, 23], naive Bayes classifiers [38, 40], causal inference [5, 25], etc. This problem has been studied under different settings and assumptions, including pairwise ($d = 2$) and mutual ($d \geq 2$) independence, univariate ($X_1, \ldots, X_d \in \mathbb{R}^1$) and multivariate ($X_1 \in \mathbb{R}^{p_1}, \ldots, X_d \in \mathbb{R}^{p_d}$) components, and more. Specifically, we focus on the general case that $X_1, \ldots, X_d$ are not assumed jointly normal.

The most extensively studied case is pairwise independence with univariate components ($X_1, X_2 \in \mathbb{R}^1$): Rank correlation is considered as a non-parametric counterpart to Pearson's product-moment correlation [28], including Kendall's τ [19], Spearman's ρ [32], etc. Bergsma and Dassios [2] proposed a test based on an extension of Kendall's τ, testing an equivalent condition to H_0. Additionally, Hoeffding [15] proposed a non-parametric test based on marginal and joint distribution functions, testing a necessary condition to investigate H_0.

[1]Research support from an NSF Award (DMS-1455172), a Xerox PARC Faculty Research Award, and Cornell University Atkinson Center for a Sustainable Future (AVF-2017).

For pairwise independence with multivariate components ($X_1 \in \mathbb{R}^{p_1}, X_2 \in \mathbb{R}^{p_2}$): Székely et al. [37], Székely and Rizzo [34] proposed a test based on distance covariance with fixed p_1, p_2 and $n \to \infty$ testing an equivalent condition to H_0, which has been extended to martingale difference divergence in Shao and Zhang [31] and Jin et al. [17] testing conditional mean independence. Under the same setting, Gretton et al. [13] proposed a test based on Hilbert−Schmidt independence criterion (HSIC), which is 0 if and only if pairwise independence holds. Further, Székely and Rizzo [35] proposed a t-test based on a modified distance covariance for the setting in which n is finite and $p_1, p_2 \to \infty$, testing an equivalent condition to H_0 as well.

For mutual independence with univariate components ($X_1, \ldots, X_d \in \mathbb{R}^1$): One natural way to extend the pairwise rank correlation to multiple components is to collect the rank correlations between all pairs of components, and examine the norm ($\mathcal{L}_2, \mathcal{L}_\infty$) of this collection. Leung and Drton [21] proposed a test based on the $\mathcal{L}_2$ norm with $n, d \to \infty$, and $d/n \to \gamma \in (0, \infty)$, and Han et al. [14] proposed a test based on the $\mathcal{L}_\infty$ norm with $n, d \to \infty$, and $d/n \to \gamma \in [0, \infty]$. Each are testing a necessary condition to H_0, in general.

For mutual independence with multivariate components ($X_1 \in \mathbb{R}^{p_1}, \ldots, X_d \in \mathbb{R}^{p_d}$): This challenging scenario has not been well studied. Using a combinatorial formula of Möbius, Genest and Rémillard [11], Genest et al. [12] and Kojadinovic and Holmes [20] proposed tests based on ranks and Cramér−von Mises statistics, testing a necessary condition to H_0; Bilodeau and Lafaye de Micheaux [3] proposed a test based on characteristic functions under the assumption of normal margins and made a connection to V-statistics, Beran et al. [1] proposed a test based on half-space probabilities, Bilodeau and Nangue [4] and Fan et al. [9] proposed tests based on characteristic functions, testing an equivalent condition to H_0, all with fixed $d, p_1, \ldots, p_d$ and $n \to \infty$. Under the same setting, Pfister et al. [30] proposed a test based on d-variable Hilbert−Schmidt independence criterion (dHSIC), which originates from HSIC and is 0 if and only if mutual independence holds. Yao et al. [39] proposed a test based on distance covariance between all pairs of components with $n, d \to \infty$, testing a necessary condition to H_0. Inspired by distance covariance in Székely et al. [37], we propose new tests based on three measures of mutual dependence, i.e., complete measure, asymmetric measure and symmetric measure, with fixed $d, p_1, \ldots, p_d$ and $n \to \infty$ in this paper, testing an equivalent condition to H_0. All computational complexities in this paper make no reference to the dimensions $d, p_1, \ldots, p_d$, as they are treated as constants.

Our measures of mutual dependence involve V-statistics, and are 0 if and only if mutual independence holds. They belong to energy statistics [36], and share many statistical properties with distance covariance. Our complete measure and dHSIC [30] both contain V-statistics with a similar structure. The main difference is that Pfister et al. [30] pursue kernel methods and overcome the computation bottleneck by resampling and Gamma approximation, while we take advantage of characteristic functions and resort to incomplete V-statistics. Our asymmetric and symmetric measures, and measures in Bilodeau and Nangue [4] and Fan et al. [9] all use characteristic functions. The main difference is that Bilodeau and Nangue [4] and Fan et al. [9] include all pairwise dependencies from the Möbius decomposition, while we only consider a subset of pairwise dependencies from it.

The weakness of testing mutual independence by a necessary condition, all pairwise independencies motivates our work on measures of mutual dependence, which is demonstrated by examples in section 6: If we directly test mutual independence based on the measures of mutual dependence proposed in this paper, we successfully detect mutual dependence. Alternatively, if we check all pairwise independencies based on distance covariance, we fail to detect any pairwise dependence, and mistakenly conclude that mutual independence holds probably because the mutual effect averages out when we narrow down to a pair.

The rest of this paper is organized as follows. In section 2, we give a brief overview of distance covariance. In section 3, we generalize distance covariance to complete measure of mutual dependence, with its properties and asymptotic distributions derived. In section 4, we propose asymmetric and symmetric measures of mutual dependence, defined as sums of squared distance covariances. We present simulation results in section 5, followed by synthetic and real data analysis in section 6[2]. Finally, section 7 is the summary of our work. All proofs have been moved to appendix.

The following notations will be used throughout this paper. Let $(\cdot, \cdot, \ldots, \cdot)$ denote a concatenation of (vector) components into a vector. Let $t = (t_1, \ldots, t_d), t^0 = (t_1^0, \ldots, t_d^0), X = (X_1, \ldots, X_d) \in \mathbb{R}^p$ where $t_j, t_j^0, X_j \in \mathbb{R}^{p_j}$, such that p_j is the marginal dimension, $j = 1, \ldots, d$, and $p = \sum_{j=1}^d p_j$ is the total dimension. The assumed "X" under H_0 is denoted by $\widetilde{X} = (\widetilde{X}_1, \ldots, \widetilde{X}_d)$, where $\widetilde{X}_j \overset{\mathcal{D}}{=} X_j$, $j = 1, \ldots, d$, $\widetilde{X}_1, \ldots, \widetilde{X}_d$ are mutually independent, and $X, \widetilde{X}$ are independent. Let X', X'' be independent copies of X, i.e., $X, X', X'' \overset{i.i.d.}{\sim} F_X$, and $\widetilde{X}', \widetilde{X}''$ be independent copies of $\widetilde{X}$, i.e., $\widetilde{X}, \widetilde{X}', \widetilde{X}'' \overset{i.i.d.}{\sim} F_{\widetilde{X}}$. The Euclidean norm of vector $X \in \mathbb{R}^p$ is denoted by $|X|_p$. Let the weighted $\mathcal{L}_2$ norm $\|\cdot\|_w$ of complex-valued function $\eta(t)$ be defined by $\|\eta(t)\|_w^2 = \int_{\mathbb{R}^p} |\eta(t)|^2 w(t)\, dt$ where $|\eta(t)|^2 = \eta(t)\overline{\eta(t)}$, $\overline{\eta(t)}$ is the complex conjugate of $\eta(t)$, and $w(t)$ is any positive weight function for which the integral exists.

Given the i.i.d. sample $\mathbf{X}$ from F_X, let $\mathbf{X}_j = \{X_j^k : k = 1, \ldots, n\}$ denote the corresponding i.i.d. sample from F_{X_j}, $j = 1, \ldots, d$, such that $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_d\}$. Denote the joint characteristic functions of X and $\widetilde{X}$ as $\phi_X(t) = \mathrm{E}[e^{i\langle t, X\rangle}]$ and $\phi_{\widetilde{X}}(t) = \prod_{j=1}^d \mathrm{E}[e^{i\langle t_j, X_j\rangle}]$, and denote the empirical versions of $\phi_X(t)$ and $\phi_{\widetilde{X}}(t)$ as $\phi_X^n(t) = \frac{1}{n}\sum_{k=1}^n e^{i\langle t, X^k\rangle}$ and $\phi_{\widetilde{X}}^n(t) = \prod_{j=1}^d (\frac{1}{n}\sum_{k=1}^n e^{i\langle t_j, X_j^k\rangle})$. For illustration purpose, we make a toy example with two components ($d = 2$), two dimensions each ($p = 4$), and two samples ($n = 2$), to exemplify the definitions of empirical measures proposed in this paper.

2. Distance Covariance

Székely et al. [37] proposed distance covariance to capture non-linear and non-monotone pairwise dependence between two random vectors ($X_1 \in \mathbb{R}^{p_1}, X_2 \in \mathbb{R}^{p_2}$).

X_1, X_2 are pairwise independent if and only if $\phi_X(t) = \phi_{X_1}(t_1)\phi_{X_2}(t_2)$, $\forall t$, which is equivalent to $\int_{\mathbb{R}^p} |\phi_X(t) - \phi_{\widetilde{X}}(t)|^2 w(t)\, dt = 0$, $\forall w(t) > 0$ if the integral exists. A class of the weight functions $w_0(t, m) = (K(p_1; m)K(p_2; m)|t_1|_{p_1}^{p_1+m}|t_2|_{p_2}^{p_2+m})^{-1}$

[2] An accompanying R package `EDMeasure` [18] is available on CRAN.

make the integral a finite and meaningful quantity composed of m-th moments according to Lemma 1 in Székely and Rizzo [33], where $K(q,m) = \frac{2\pi^{q/2}\Gamma(1-m/2)}{m2^m\Gamma((q+m)/2)}$, and Γ is the gamma function.

The non-negative distance covariance $\mathcal{V}(X)$ is defined by $\mathcal{V}^2(X) = \|\phi_X(t)-\phi_{\widetilde{X}}(t)\|^2_{w_0} = \int_{\mathbb{R}^p} |\phi_X(t)-\phi_{\widetilde{X}}(t)|^2 w_0(t)\,dt$, where

$$w_0(t) = (K_{p_1}K_{p_2}|t_1|_{p_1}^{p_1+1}|t_2|_{p_2}^{p_2+1})^{-1},$$

with $m = 1$ and $K_q = K(q,1)$, while any following result can be generalized to $0 < m < 2$. If $\mathrm{E}|X|_p < \infty$, then $\mathcal{V}(X) \in [0,\infty)$, and $\mathcal{V}(X) = 0$ if and only if X_1, X_2 are pairwise independent.

The non-negative empirical distance covariance $\mathcal{V}_n(\mathbf{X})$ is defined by $\mathcal{V}_n^2(\mathbf{X}) = \|\phi_X^n(t) - \phi_{\widetilde{X}}^n(t)\|^2_{w_0} = \int_{\mathbb{R}^p} |\phi_X^n(t) - \phi_{\widetilde{X}}^n(t)|^2 w_0(t)\,dt$. Calculating $\mathcal{V}_n^2(\mathbf{X})$ via the symmetry of Euclidian distances has the time complexity $O(n^2)$. Some asymptotic properties of $\mathcal{V}_n(\mathbf{X})$ are derived. If $\mathrm{E}|X|_p < \infty$, then (i) $\mathcal{V}_n(\mathbf{X}) \xrightarrow[n\to\infty]{a.s.} \mathcal{V}(X)$. (ii) Under H_0, $n\mathcal{V}_n^2(\mathbf{X}) \xrightarrow[n\to\infty]{\mathcal{D}} \|\zeta(t)\|^2_{w_0}$ where $\zeta(t)$ is a complex-valued Gaussian process with mean zero and covariance function $R(t,t^0) = [\phi_{X_1}(t_1 - t_1^0) - \phi_{X_1}(t_1)\overline{\phi_{X_1}(t_1^0)}][\phi_{X_2}(t_2 - t_2^0) - \phi_{X_2}(t_2)\overline{\phi_{X_2}(t_2^0)}]$. (iii) Under H_A, $n\mathcal{V}_n^2(\mathbf{X}) \xrightarrow[n\to\infty]{a.s.} \infty$.

3. Complete Measure of Mutual Dependence

Generalizing the idea of distance covariance, we propose complete measure of mutual dependence to capture non-linear and non-monotone mutual dependence between multiple random vectors ($X_1 \in \mathbb{R}^{p_1}, \ldots, X_d \in \mathbb{R}^{p_d}$).

$X_1, \ldots, X_d$ are mutually independent if and only if $\phi_X(t) = \phi_{X_1}(t_1)\ldots\phi_{X_d}(t_d) = \phi_{\widetilde{X}}(t)$, $\forall t$, which is equivalent to $\int_{\mathbb{R}^p} |\phi_X(t) - \phi_{\widetilde{X}}(t)|^2 w(t)\,dt = 0$, $\forall w(t) > 0$ if the integral exists. In the following, we will present two weights $w_1(t), w_2(t)$, and elaborate on the reason why we disregard $w_2(t)$ for computational efficiency later.

We put all components together instead of separating them, and choose the weight function

$$w_1(t) = (K_p|t|_p^{p+1})^{-1}.$$

Definition 1. The complete measure of mutual dependence $\mathcal{Q}(X)$ is defined by

$$\mathcal{Q}(X) = \|\phi_X(t) - \phi_{\widetilde{X}}(t)\|^2_{w_1} = \int_{\mathbb{R}^p} |\phi_X(t) - \phi_{\widetilde{X}}(t)|^2 w_1(t)\,dt.$$

We can show an equivalence to mutual independence based on $\mathcal{Q}(X)$ according to Lemma 1 in Székely and Rizzo [33].

Theorem 1. *If $E|X|_p < \infty$, then $\mathcal{Q}(X) \in [0,\infty)$, and $\mathcal{Q}(X) = 0$ if and only if $X_1, \ldots, X_d$ are mutually independent. In addition, $\mathcal{Q}(X)$ has an interpretation as expectations*

$$\mathcal{Q}(X) = E|X - \widetilde{X}'|_p + E|X' - \widetilde{X}|_p - E|X - X'|_p - E|\widetilde{X} - \widetilde{X}'|_p.$$

It is straightforward to estimate $\mathcal{Q}(X)$ by replacing the characteristic functions with the empirical characteristic functions from the sample.

Definition 2. The empirical complete measure of mutual dependence $\mathcal{Q}_n(\mathbf{X})$ is defined by

$$\mathcal{Q}_n(\mathbf{X}) = \|\phi_X^n(t) - \phi_{\widetilde{X}}^n(t)\|^2_{w_1} = \int_{\mathbb{R}^p} |\phi_X^n(t) - \phi_{\widetilde{X}}^n(t)|^2 w_1(t)\,dt.$$

Lemma 1. *$Q_n(\mathbf{X})$ has an interpretation as complete V-statistics*

$$\begin{aligned} Q_n(\mathbf{X}) &= \frac{2}{n^{d+1}} \sum_{k,\ell_1,\ldots,\ell_d=1}^{n} |X^k - (X_1^{\ell_1}, \ldots, X_d^{\ell_d})|_p + \frac{1}{n^2} \sum_{k,\ell=1}^{n} |X^k - X^\ell|_p \\ &\quad - \frac{1}{n^{2d}} \sum_{k_1,\ldots,k_d,\ell_1,\ldots,\ell_d=1}^{n} |(X_1^{k_1}, \ldots, X_d^{k_d}) - (X_1^{\ell_1}, \ldots, X_d^{\ell_d})|_p, \end{aligned}$$

whose naive implementation has the time complexity $O(n^{2d})$.

With respect to the toy example, the first summation term in $Q_n(\mathbf{X})$ contains 8 summands $|(X_1^k - X_1^{\ell_1}, X_2^k - X_2^{\ell_2})|_4$, $\forall k, \ell_1, \ell_2 \in \{1, 2\}$, including $|(X_1^1 - X_1^1, X_2^1 - X_2^2)|_4$ and $|(X_1^1 - X_1^2, X_2^1 - X_2^2)|_4$.

In view of the definition of distance covariance, it may seem natural to define the measure using the weight function

$$w_2(t) = (K_{p_1} \ldots K_{p_d} |t_1|_{p_1}^{p_1+1} \ldots |t_d|_{p_d}^{p_d+1})^{-1},$$

which equals $w_0(t)$ when $d = 2$. Given the weight function $w_2(t)$, we can define the squared distance covariance of mutual dependence $\mathcal{U}(X) = \|\phi_X(t) - \phi_{\overline{X}}(t)\|_{w_2}^2$ and its empirical counterpart $\mathcal{U}_n(\mathbf{X}) = \|\phi_X^n(t) - \phi_{\overline{X}}^n(t)\|_{w_2}^2$, which equal $\mathcal{V}^2(X)$ and $\mathcal{V}_n^2(\mathbf{X})$ when $d = 2$. The naive implementation of $\mathcal{U}_n(\mathbf{X})$ has the time complexity $O(n^{d+1})$.

The reason to favor $w_1(t)$ instead of $w_2(t)$ is a trade-off between the moment condition and time complexity. We often cannot afford the time complexity of $Q_n(\mathbf{X})$ or $\mathcal{U}_n(\mathbf{X})$, and have to simplify them through incomplete V-statistics. An incomplete V-statistic is obtained by sampling the terms of a complete V-statistic, where the summation extends over only a subset of the tuple of indices. To simplify by replacing complete V-statistics with incomplete V-statistics, $\mathcal{U}_n(\mathbf{X})$ requires the additional d-th moment condition $\mathrm{E}(|X_1|_{p_1} \ldots |X_d|_{p_d}) < \infty$, while $Q_n(\mathbf{X})$ does not require any other condition in addition to the first moment condition $\mathrm{E}|X|_p < \infty$. Thus, we can reduce the complexity of $Q_n(\mathbf{X})$ to $O(n^2)$ with a weaker condition, which makes $Q(X)$ and $Q_n(\mathbf{X})$ from $w_1(t)$ a more general solution. As an example, suppose $X_1 = \ldots = X_d \in \mathbb{R}^1$, then $\mathrm{E}|X|_p < \infty$ only requires finite first moment as $\mathrm{E}|X_1| < \infty$, while $\mathrm{E}(|X_1|_{p_1} \ldots |X_d|_{p_d}) < \infty$ requires finite d-th moment as $\mathrm{E}|X_1|^d < \infty$.

Moreover, we define the simplified empirical version of $\phi_{\overline{X}}(t)$ as

$$\phi_{\overline{X}}^{n\star}(t) = \frac{1}{n} \sum_{k=1}^{n} e^{i \sum_{j=1}^{d} \langle t_j, X_j^{k+j-1} \rangle} = \frac{1}{n} \sum_{k=1}^{n} e^{i\langle t, (X_1^k, \ldots, X_d^{k+d-1}) \rangle},$$

in order to substitute $\phi_{\overline{X}}^n(t)$ for simplification, where X_j^{n+k} is interpreted as X_j^k for $k > 0$.

Definition 3. The simplified empirical complete measure of mutual dependence $Q_n^\star(\mathbf{X})$ is defined by

$$Q_n^\star(\mathbf{X}) = \|\phi_X^n(t) - \phi_{\overline{X}}^{n\star}(t)\|_{w_1}^2 = \int_{\mathbb{R}^p} |\phi_X^n(t) - \phi_{\overline{X}}^{n\star}(t)|^2 w_1(t)\, dt.$$

Lemma 2. *$Q_n^\star(\mathbf{X})$ has an interpretation as incomplete V-statistics*

$$\begin{aligned} Q_n^\star(\mathbf{X}) &= \frac{2}{n^2}\sum_{k,\ell=1}^{n}|X^k-(X_1^\ell,\ldots,X_d^{\ell+d-1})|_p+\frac{1}{n^2}\sum_{k,\ell=1}^{n}|X^k-X^\ell|_p \\ &\quad -\frac{1}{n^2}\sum_{k,\ell=1}^{n}|(X_1^k,\ldots,X_d^{k+d-1})-(X_1^\ell,\ldots,X_d^{\ell+d-1})|_p, \end{aligned}$$

whose naive implementation has the time complexity $O(n^2)$.

With respect to the toy example, the first summation term in $Q_n^\star(\mathbf{X})$ contains 4 summands $|(X_1^k-X_1^\ell, X_2^k-X_2^{\ell+1})|_4$, $\forall k,\ell\in\{1,2\}$, including $|(X_1^1-X_1^1, X_2^1-X_2^2)|_4$ but not $|(X_1^1-X_1^2, X_2^1-X_2^2)|_4$.

Using a similar derivation to Theorem 2 and 5 of Székely et al. [37], some asymptotic distributions of $Q_n(\mathbf{X}), Q_n^\star(\mathbf{X})$ are obtained as follows.

Theorem 2. *If $E|X|_p<\infty$, then*

$$Q_n(\mathbf{X})\xrightarrow[n\to\infty]{a.s.}Q(X)\quad\textit{and}\quad Q_n^\star(\mathbf{X})\xrightarrow[n\to\infty]{a.s.}Q(X).$$

Theorem 3. *If $E|X|_p<\infty$, then under H_0, we have*

$$nQ_n(\mathbf{X})\xrightarrow[n\to\infty]{\mathcal{D}}\|\zeta(t)\|_{w_1}^2\quad\textit{and}\quad nQ_n^\star(\mathbf{X})\xrightarrow[n\to\infty]{\mathcal{D}}\|\zeta^\star(t)\|_{w_1}^2,$$

where $\zeta(t),\zeta^\star(t)$ are complex-valued Gaussian processes with mean zero and covariance functions

$$\begin{aligned} R(t,t^0) &= \prod_{j=1}^{d}\phi_{X_j}(t_j-t_j^0)+(d-1)\prod_{j=1}^{d}\phi_{X_j}(t_j)\overline{\phi_{X_j}(t_j^0)}-\sum_{j=1}^{d}\phi_{X_j}(t_j-t_j^0)\prod_{\ell\neq j}\phi_{X_\ell}(t_\ell)\overline{\phi_{X_\ell}(t_\ell^0)}, \\ R^\star(t,t^0) &= 2R(t,t^0). \end{aligned}$$

Under H_A, we have

$$nQ_n(\mathbf{X})\xrightarrow[n\to\infty]{a.s.}\infty\quad\textit{and}\quad nQ_n^\star(\mathbf{X})\xrightarrow[n\to\infty]{a.s.}\infty.$$

Theorem 2 and 3 are closely connected in the sense that $nQ_n(\mathbf{X}), nQ_n^\star(\mathbf{X})$ diverges to ∞ under H_A as $Q_n(\mathbf{X}), Q_n^\star(\mathbf{X})$ converges to $Q(\mathbf{X}), Q^\star(\mathbf{X})$. Furthermore, $nQ_n(\mathbf{X}), nQ_n^\star(\mathbf{X})$ converges to a proper random variable under H_0, which implies $Q_n(\mathbf{X}), Q_n^\star(\mathbf{X})$ converges to 0 under H_0.

Therefore, a mutual independence test can be proposed based on the weak convergence of $nQ_n(\mathbf{X}), nQ_n^\star(\mathbf{X})$ in Theorem 3. Since the asymptotic null distributions of $nQ_n(\mathbf{X}), nQ_n^\star(\mathbf{X})$ depend on F_X, they will not be used in practice, and a permutation procedure will be used to approximate them instead.

4. Asymmetric and Symmetric Measures of Mutual Dependence

As an alternative, we now propose the asymmetric and symmetric measures of mutual dependence to capture mutual dependence via aggregating pairwise dependencies.

The subset of components on the right of X_c is denoted by $X_{c^+}=(X_{c+1},\ldots,X_d)$, with $t_{c^+}=(t_{c+1},\ldots,t_d)$, $c=0,1,\ldots,d-1$. The subset of components except X_c is denoted by $X_{-c}=(X_1,\ldots,X_{c-1},X_{c^+})$, with $t_{-c}=(t_1,\ldots,t_{c-1},t_{c^+})$, $c=1,\ldots,d-1$.

We denote pairwise independence by $\perp$. The collection of pairwise independencies implied by mutual independence includes "one versus others on the right"

$$\{X_1 \perp X_{1^+}, X_2 \perp X_{2^+}, \ldots, X_{d-1} \perp X_d\}, \tag{1}$$

"one versus all the others"

$$\{X_1 \perp X_{-1}, X_2 \perp X_{-2}, \ldots, X_d \perp X_{-d}\}, \tag{2}$$

and many others, e.g., $(X_1, X_2) \perp X_{2^+}$. In fact, the number of pairwise independencies resulting from mutual independence is at least $2^{d-1} - 1$, which grows exponentially with the number of components d. Therefore, we cannot test mutual independence simply by checking all pairwise independencies even with moderate d.

Fortunately, we have two options to test only a small subset of all pairwise independencies to fulfill the task. The first one is that H_0 holds if and only if (1) holds, which can be verified via the sequential decomposition of distribution functions. This option is asymmetric and not unique, having $d!$ feasible subsets with respect to different orders of $X_1, \ldots, X_d$. The second one is that H_0 holds if and only if (2) holds, which can be verified via the stepwise decomposition of distribution functions and the fact that $X_j \perp X_{-j}$ implies $X_j \perp X_{j^+}$. This option is symmetric and unique, having only one feasible subset.

To shed light on why these two options are necessary and sufficient conditions to mutual independence, we present the following inequality that the mutual dependence can be bounded by a sum of several pairwise dependencies as

$$|\phi_X(t) - \prod_{j=1}^{d} \phi_{X_j}(t_j)| \leq \sum_{c=1}^{d-1} |\phi_{(X_c, X_{c^+})}((t_c, t_{c^+})) - \phi_{X_c}(t_c)\phi_{X_{c^+}}(t_{c^+})|^2.$$

In consideration of these two options, we test a set of pairwise independencies in place of mutual independence, where we use $\mathcal{V}^2(X)$ to test pairwise independence.

Definition 4. The asymmetric and symmetric measures of mutual dependence $\mathcal{R}(X), \mathcal{S}(X)$ are defined by

$$\mathcal{R}(X) = \sum_{c=1}^{d-1} \mathcal{V}^2((X_c, X_{c^+})) \quad \textit{and} \quad \mathcal{S}(X) = \sum_{c=1}^{d} \mathcal{V}^2((X_c, X_{-c})).$$

We can show an equivalence to mutual independence based on $\mathcal{R}(X), \mathcal{S}(X)$ according to Theorem 3 of Székely et al. [37].

Theorem 4. *If $E|X|_p < \infty$, then $\mathcal{R}(X), \mathcal{S}(X) \in [0, \infty)$, and $\mathcal{R}(X), \mathcal{S}(X) = 0$ if and only if $X_1, \ldots, X_d$ are mutually independent.*

It is straightforward to estimate $\mathcal{R}(X), \mathcal{S}(X)$ by replacing the characteristic functions with the empirical characteristic functions from the sample.

Definition 5. The empirical asymmetric and symmetric measures of mutual dependence $\mathcal{R}_n(\mathbf{X}), \mathcal{S}_n(\mathbf{X})$ are defined by

$$\mathcal{R}_n(\mathbf{X}) = \sum_{c=1}^{d-1} \mathcal{V}_n^2((\mathbf{X}_c, \mathbf{X}_{c^+})) \quad \textit{and} \quad \mathcal{S}_n(\mathbf{X}) = \sum_{c=1}^{d} \mathcal{V}_n^2((\mathbf{X}_c, \mathbf{X}_{-c})).$$

The implementations of $\mathcal{R}_n(\mathbf{X}), \mathcal{S}_n(\mathbf{X})$ have the time complexity $O(n^2)$. Using a similar derivation to Theorem 2 and 5 of Székely et al. [37], some asymptotic properties of $\mathcal{R}_n(\mathbf{X}), \mathcal{S}_n(\mathbf{X})$ are obtained as follows.

Theorem 5. *If $E|X|_p < \infty$, then*

$$\mathcal{R}_n(\mathbf{X}) \xrightarrow[n\to\infty]{a.s.} \mathcal{R}(X) \quad \textit{and} \quad \mathcal{S}_n(\mathbf{X}) \xrightarrow[n\to\infty]{a.s.} \mathcal{S}(X).$$

Theorem 6. *If $E|X|_p < \infty$, then under H_0, we have*

$$n\mathcal{R}_n(\mathbf{X}) \xrightarrow[n\to\infty]{\mathcal{D}} \sum_{j=1}^{d-1} \|\zeta_j^R((t_j, t_{j^+}))\|_{w_0}^2 \quad \textit{and} \quad n\mathcal{S}_n(\mathbf{X}) \xrightarrow[n\to\infty]{\mathcal{D}} \sum_{j=1}^{d} \|\zeta_j^S((t_j, t_{-j}))\|_{w_0}^2,$$

where $\zeta_j^R((t_j, t_{j^+})), \zeta_j^S((t_j, t_{-j}))$ are complex-valued Gaussian processes corresponding to the limiting distributions of $n\mathcal{V}_n^2((\mathbf{X}_j, \mathbf{X}_{j^+})), n\mathcal{V}_n^2((\mathbf{X}_j, \mathbf{X}_{-j}))$. Under H_A, we have

$$n\mathcal{R}_n(\mathbf{X}) \xrightarrow[n\to\infty]{a.s.} \infty \quad \textit{and} \quad n\mathcal{S}_n(\mathbf{X}) \xrightarrow[n\to\infty]{a.s.} \infty.$$

It is surprising to find that $\mathcal{V}_n^2((\mathbf{X}_c, \mathbf{X}_{c^+})), c = 1, \ldots, d-1$ are mutually independent asymptotically, and $\mathcal{V}_n^2((\mathbf{X}_c, \mathbf{X}_{-c})), c = 1, \ldots, d$ are mutually independent asymptotically as well, which is a crucial discovery behind Theorem 6. Theorem 5 and 6 are also closely connected in a similar way to Theorem 2 and 3. Similar to Theorem 3, the asymptotic results in Theorem 6 will not be used, but will be approximated by a permutation procedure in the tests.

$\mathcal{R}_n(\mathbf{X}), \mathcal{S}_n(\mathbf{X})$ only contain a subset of pairwise dependencies from the Möbius decomposition used in Bilodeau and Nangue [4] and Fan et al. [9], but we still obtain an equivalent condition to mutual independence. On the one hand, $\mathcal{R}_n(\mathbf{X}), \mathcal{S}_n(\mathbf{X})$ have much lower complexity when d gets large. On the other hand, we probably cannot narrow down to the smallest pair with significant dependence, while we can still find clues about the dependence structure. For example, the dependence between X_1 and X_2 is not directly included in $\mathcal{R}_n(\mathbf{X}), \mathcal{S}_n(\mathbf{X})$, but it is expected to be captured by the dependence between X_1 and X_{1^+} included in $\mathcal{R}_n(\mathbf{X}), \mathcal{S}_n(\mathbf{X})$. Thus, we can observe the dependence between X_1 and X_{1^+}, but not between X_1 and X_2 without further investigation.

Alternatively, we can plug in $\mathcal{Q}(X)$ instead of $\mathcal{V}^2(X)$ in Definition 4 and $\mathcal{Q}_n(\mathbf{X})$ instead of $\mathcal{V}_n^2(\mathbf{X})$ in Definition 5, and define the asymmetric and symmetric measures $\mathcal{J}(X), \mathcal{I}(X)$ accordingly, which equal $\mathcal{Q}(X), \mathcal{Q}_n(\mathbf{X})$ when $d = 2$. The naive implementations of $\mathcal{J}_n(\mathbf{X}), \mathcal{I}_n(\mathbf{X})$ have the time complexity $O(n^4)$. Similarly, we can replace $\mathcal{Q}_n(\mathbf{X})$ with $\mathcal{Q}_n^\star(\mathbf{X})$ to simplify them, and define the simplified empirical asymmetric and symmetric measures $\mathcal{J}_n^\star(\mathbf{X}), \mathcal{I}_n^\star(\mathbf{X})$, reducing their complexities to $O(n^2)$ without any other condition except the first moment condition $\mathrm{E}|X|_p < \infty$. Through the same derivations, we can show that $\mathcal{J}_n(\mathbf{X}), \mathcal{J}_n^\star(\mathbf{X})$, $\mathcal{I}_n(\mathbf{X}), \mathcal{I}_n^\star(\mathbf{X})$ have similar convergences as $\mathcal{R}_n(\mathbf{X}), \mathcal{S}_n(\mathbf{X})$ in Theorem 5 and 6.

5. Simulation Studies

In this section, we evaluate the finite sample performance of proposed measures $\mathcal{Q}_n, \mathcal{R}_n, \mathcal{S}_n, \mathcal{J}_n, \mathcal{I}_n, \mathcal{Q}_n^\star, \mathcal{J}_n^\star, \mathcal{I}_n^\star$ by performing simulations similar to Székely et al. [37], and compare them with benchmark measures $\mathcal{V}_n^2$ [37],

BN^h, BN^d [4], dHSIC [30], and HL^τ, HL^ρ [14] respectively in various scenarios. Note that BN^h is based on HSIC, BN^d is based on distance covariance, HL^τ is based on Kendall's τ, and HL^ρ is based on Spearman's ρ. We also include permutation tests based on finite-sample extensions of $\mathrm{HL}^\tau, \mathrm{HL}^\rho$, denoted by $\mathrm{HL}^\tau_n, \mathrm{HL}^\rho_n$. Moreover, dHSIC is implemented in the R package `dHSIC` [29] using the gaussian kernel with a median heuristic to choose the bandwidth.

We test the null hypothesis H_0 with significance level $\alpha = 0.1$ and examine the empirical size and power of each measure. In each scenario, we run 1,000 repetitions with the adaptive permutation size $B = \lfloor 200 + 5000/n \rfloor$ where n is the sample size, for all empirical measures that require a permutation procedure to approximate their asymptotic distributions, i.e., $\mathcal{Q}_n, \mathcal{R}_n, \mathcal{S}_n, \mathcal{J}_n, \mathcal{I}_n, \mathcal{Q}_n^\star, \mathcal{J}_n^\star, \mathcal{I}_n^\star, \mathcal{V}_n^2, \mathrm{BN}^h, \mathrm{BN}^d, \mathrm{dHSIC}, \mathrm{HL}^\tau_n, \mathrm{HL}^\rho_n$.

In the following two examples, we fix $d = 2$ and change n from 25 to 500, and compare $\mathcal{Q}_n, \mathcal{R}_n, \mathcal{S}_n, \mathcal{J}_n, \mathcal{I}_n, \mathcal{Q}_n^\star, \mathcal{J}_n^\star, \mathcal{I}_n^\star$ to $\mathcal{V}_n^2$.

Example 1 (pairwise multivariate normal)**.** $X_1, X_2 \in \mathbb{R}^5$, $(X_1, X_2)^\top \sim \mathcal{N}_{10}(0, \Sigma)$ where $\Sigma_{ii} = 1$. Under H_0, $\Sigma_{ij} = 0$, $i \neq j$. Under H_A, $\Sigma_{ij} = 0.1$, $i \neq j$. See results in Table 1 and 2.

Example 2 (pairwise multivariate non-normal)**.** $X_1, X_2 \in \mathbb{R}^5$, $(Y_1, Y_2)^\top \sim \mathcal{N}_{10}(0, \Sigma)$ where $\Sigma_{ii} = 1$. $X_1 = \ln(Y_1^2), X_2 = \ln(Y_2^2)$. Under H_0, $\Sigma_{ij} = 0$, $i \neq j$. Under H_A, $\Sigma_{ij} = 0.4$, $i \neq j$. See results in Table 3 and 4.

For both Example 1 and 2, the empirical size of all measures is close to $\alpha = 0.1$. The empirical power of $\mathcal{Q}_n, \mathcal{R}_n, \mathcal{S}_n, \mathcal{J}_n, \mathcal{I}_n$ is almost the same as that of $\mathcal{V}_n^2$, while the empirical power of $\mathcal{Q}_n^\star, \mathcal{J}_n^\star, \mathcal{I}_n^\star$ is lower than that of $\mathcal{V}_n^2$, which makes sense because we trade-off testing power and time complexity for simplified measures.

In the following two examples, we fix $d = 3$ and change n from 25 to 500, and compare $\mathcal{Q}_n, \mathcal{R}_n, \mathcal{S}_n, \mathcal{J}_n, \mathcal{I}_n, \mathcal{Q}_n^\star, \mathcal{J}_n^\star, \mathcal{I}_n^\star$ to $\mathrm{BN}^h, \mathrm{BN}^d, \mathrm{dHSIC}$.

Example 3 (mutual multivariate normal)**.** $X_1, X_2, X_3 \in \mathbb{R}^5$, $(X_1, X_2, X_3)^\top \sim \mathcal{N}_{15}(0, \Sigma)$ where $\Sigma_{ii} = 1$. Under H_0, $\Sigma_{ij} = 0$, $i \neq j$. Under H_A, $\Sigma_{ij} = 0.1$, $i \neq j$. See results in Table 5 and 6.

Example 4 (mutual multivariate non-normal)**.** $X_1, X_2, X_3 \in \mathbb{R}^5$. $(Y_1, Y_2, Y_3)^\top \sim \mathcal{N}_{15}(0, \Sigma)$ where $\Sigma_{ii} = 1$. $X_k = \ln(Y_k^2)$, $k = 1, 2, 3$. Under H_0, $\Sigma_{ij} = 0$, $i \neq j$. Under H_A, $\Sigma_{ij} = 0.4$, $i \neq j$. See results in Table 7 and 8.

For both Example 3 and 4, the empirical size of all measures is close to $\alpha = 0.1$. The empirical power of $\mathcal{Q}_n, \mathcal{R}_n, \mathcal{S}_n, \mathcal{J}_n, \mathcal{I}_n$ is almost the same, the empirical power of $\mathcal{Q}_n^\star, \mathcal{J}_n^\star, \mathcal{I}_n^\star$ is almost the same, while the empirical power of $\mathcal{Q}_n^\star, \mathcal{J}_n^\star, \mathcal{I}_n^\star$ is lower than that of $\mathcal{Q}_n, \mathcal{R}_n, \mathcal{S}_n, \mathcal{J}_n, \mathcal{I}_n$, which makes sense since we trade-off testing power and time complexity for simplified measures. $\mathrm{BN}^h, \mathrm{BN}^d, \mathrm{dHSIC}$ outperform all other measures in the normal Example 3, while $\mathcal{Q}_n, \mathcal{R}_n, \mathcal{S}_n, \mathcal{J}_n, \mathcal{I}_n$ achieves slightly better performance than $\mathrm{BN}^h, \mathrm{BN}^d, \mathrm{dHSIC}$ in the non-normal Example 4.

To compare the computation time of these measures, we evaluate one case in Example 4 with $n = 25$ under H_0. When running on Dell PowerEdge 2650 with 16GB RAM using a single core, $\mathcal{R}_n, \mathcal{S}_n, \mathcal{Q}_n^\star, \mathcal{J}_n^\star, \mathcal{I}_n^\star$ takes 164.09, 117.57, 51.66, 71.39, 94.96 seconds respectively, while $\mathrm{BN}^h, \mathrm{BN}^d, \mathrm{dHSIC}$ takes 207.16, 204.42, 70.40 seconds respectively.

In the last example, we change d from 5 to 50 and fix $n = 100$, and compare $\mathcal{R}_n, \mathcal{S}_n, \mathcal{Q}_n^\star, \mathcal{J}_n^\star, \mathcal{I}_n^\star$ to $\mathrm{HL}^\tau, \mathrm{HL}^\rho, \mathrm{HL}^\tau_n, \mathrm{HL}^\rho_n$.

Example 5 (mutual univariate normal high-dimensional)**.** $X_1, \ldots, X_d \in \mathbb{R}^1$. $(X_1, \ldots, X_d)^\top \sim \mathcal{N}_d(0, \Sigma)$ where $\Sigma_{ii} = 1$. Under H_0, $\Sigma_{ij} = 0$, $i \neq j$. Under H_A, $\Sigma_{ij} = 0.1$, $i \neq j$. See results in Table 9 and 10.

The empirical size of $\mathrm{HL}^\tau, \mathrm{HL}^\rho$ is much lower than $\alpha = 0.1$ and too conservative, while that of other measures is fairly close to $\alpha = 0.1$. The reason is probably that the convergence to asymptotic distributions of $\mathrm{HL}^\tau, \mathrm{HL}^\rho$ requires larger sample size n and number of components d. The measures $\mathcal{R}_n, \mathcal{S}_n$ have the highest empirical power, and outperform the simplified measures $\mathcal{Q}_n^\star, \mathcal{J}_n^\star, \mathcal{I}_n^\star$. The empirical power of simplified measures is similar to or even lower than that of benchmark measures when $d = 5$. However, the empirical power of simplified measures converges much faster than that of benchmark measures as d grows.

Moreover, $\mathcal{Q}_n^\star$ shows significant advantage over $\mathcal{J}_n^\star, \mathcal{I}_n^\star$. The reason is probably that $\mathcal{Q}_n^\star$ is based on truly mutual dependence while $\mathcal{J}_n^\star, \mathcal{I}_n^\star$ is based on pairwise dependencies, and large d compared to n introduces much more noise to $\mathcal{J}_n^\star, \mathcal{I}_n^\star$ because their summation structures, which makes them more difficult to detect mutual dependence.

The asymptotic analysis of our measures only allows small d compared to n, while our measures work well with large d compared to n in Example 5. However, this success relies on the underlying dependence structure, which is dense since each component is dependent on any other component. In contrast, if the dependence structure is sparse as each component is dependent on only a few of other components, then all measures are likely to fail.

6. Illustrative Examples

We start with two examples comparing different methods to show the value of our mutual independence tests. In practice, people usually check all pairwise dependencies to test mutual independence, due to the lack of reliable and universal mutual independence tests. It is very likely to miss the complicated mutual dependence structure, and make unsound decisions in corresponding applications assuming that mutual independence holds.

6.1. Synthetic Data

We define a triplet of random vectors (X, Y, Z) on $\mathbb{R}^q \times \mathbb{R}^q \times \mathbb{R}^q$, where $X, Y \sim \mathcal{N}(0, I_q)$, $W \sim \mathrm{Exp}(1/\sqrt{2})$, the first element of Z is $Z_1 = \mathrm{sign}(X_1 Y_1)W$ and the remaining $q-1$ elements are $Z_{2:q} \sim \mathcal{N}(0, I_{q-1})$, and $X, Y, W, Z_{2:q}$ are mutually independent. Clearly, (X, Y, Z) is a pairwise independent but mutually dependent triplet.

An i.i.d. sample of (X, Y, Z) is randomly generated with sample size $n = 500$ and dimension $q = 5$. On the one hand, we test the null hypothesis $H_0 : X, Y, Z$ are mutually independent using proposed measures $\mathcal{R}_n, \mathcal{S}_n, \mathcal{Q}_n^\star, \mathcal{J}_n^\star, \mathcal{I}_n^\star$. On the other hand, we test the null hypotheses $H_0^{(1)} : X \perp\!\!\!\perp Y$, $H_0^{(2)} : Y \perp\!\!\!\perp Z$, and $H_0^{(3)} : X \perp\!\!\!\perp Z$ using distance covariance $\mathcal{V}_n^2$. An adaptive permutation size $B = 210$ is used for all tests.

As expected, mutual dependence is successfully captured, as the p-values of mutual independence tests are 0.0143 ($\mathcal{Q}_n^\star$), 0.0286 ($\mathcal{J}_n^\star$), 0 ($\mathcal{I}_n^\star$), 0.0381 ($\mathcal{R}_n$) and 0 ($\mathcal{S}_n$). Meanwhile, the p-values of pairwise independence tests are 0.2905 (X, Y), 0.2619 (Y, Z), and 0.3048 (X, Z). According to the Bonferroni correction for multiple tests

among all the pairs, the significance level should be adjusted as $\alpha/3$ for pairwise tests. As a result, no signal of pairwise dependence is detected, and we cannot reject mutual independence.

6.2. Financial Data

Fama and French [6] and Fama and French [7] proposed the Fama−French three-factor and five-factor models to explain the stock returns, and demonstrated that these factors comprising the stock returns are correlated according to long-term market research in finance. Thus, we apply our tests to a subset of these factors and confirm this argument as an application.

We collect the annual Fama−French 5 factors in the past 52 years between 1964 and 2015[3]. In particular, we are interested in whether mutual dependence among three factors, X = Mkt-RF (excess return on the market), Y = SMB (small minus big), and Z = RF (risk-free return) exists, where annual returns are considered as nearly independent observations. Both histograms and pair plots of X, Y, Z are depicted in Figure 1.

For one, we apply a single mutual independence test $H_0 : X, Y, Z$ are mutually independent. For another, we apply three pairwise independence tests $H_0^{(1)} : X \perp\!\!\!\perp Y$, $H_0^{(2)} : Y \perp\!\!\!\perp Z$, and $H_0^{(3)} : X \perp\!\!\!\perp Z$. An adaptive permutation size $B = 296$ is used for all tests.

The p-values of mutual independence tests are 0.0236 ($\mathcal{Q}_n^\star$), 0.0642 ($\mathcal{J}_n^\star$), 0.0541 ($\mathcal{I}_n^\star$), 0.1588 ($\mathcal{R}_n$) and 0.1486 ($\mathcal{S}_n$), indicating that mutual dependence is successfully captured. In the meanwhile, the p-values of pairwise independence tests using distance covariance $\mathcal{V}_n^2$ are 0.1419 (X, Y), 0.5743 (Y, Z) and 0.5405 (X, Z). Similarly, the significance level should be adjusted as $\alpha/3$ according to the Bonferroni correction, and thus we cannot reject mutual independence, since no signal of pairwise dependence is detected.

7. Conclusion

We propose three measures of mutual dependence for random vectors based on the equivalence to mutual independence through characteristic functions, following the idea of distance covariance in Székely et al. [37].

When we select the weight function for the complete measure, we trade off between moment condition and time complexity. Then we simplify it by replacing complete V-statistics by incomplete V-statistics, as a trade-off between testing power and time complexity. These two trade-offs make the simplified complete measure both effective and efficient.

The asymptotic distributions of our measures depend on the underlying distribution F_X. Thus, the corresponding tests are not distribution-free, and we use a permutation procedure to approximate the asymptotic distributions in practice.

We illustrate the value of our measures through both synthetic and financial data examples, where mutual independence tests based on our measures successfully capture the mutual dependence, while the alternative checking all pairwise independencies fails and mistakenly leads to the conclusion that mutual independence holds. Our

[3]Data at http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

measures achieve competitive or even better results than the benchmark measures in simulations with various examples. Although we do not allow large d compared to n in asymptotic analysis, our measures work well in a large d example since the dependence structure is dense. Lastly, it would be interesting to extend current results on continuous variables to categorical variables, as applied statisticians may rely on such measures to conduct sensitivity analyses [24] correspondingly.

Acknowledgements

We are grateful to Stanislav Volgushev for helpful comments on a preliminary draft of this paper.

[1] R. Beran, M. Bilodeau, and P. Lafaye de Micheaux. Nonparametric tests of independence between random vectors. *Journal of Multivariate Analysis*, 98(9):1805–1824, 2007.

[2] W. Bergsma and A. Dassios. A consistent test of independence based on a sign covariance related to kendall's tau. *Bernoulli*, 20(2): 1006–1028, 2014.

[3] M. Bilodeau and P. Lafaye de Micheaux. A multivariate empirical characteristic function test of independence with normal marginals. *Journal of Multivariate Analysis*, 95(2):345–369, 2005.

[4] M. Bilodeau and A. G. Nangue. Tests of mutual or serial independence of random vectors with applications. *The Journal of Machine Learning Research*, 18(1):2518–2557, 2017.

[5] P. Ding et al. A paradox from randomization-based causal inference. *Statistical science*, 32(3):331–345, 2017.

[6] E. F. Fama and K. R. French. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56, 1993.

[7] E. F. Fama and K. R. French. A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22, 2015.

[8] J. Fan, Y. Feng, and L. Xia. A conditional dependence measure with applications to undirected graphical models. *arXiv preprint arXiv:1501.01617*, 2015.

[9] Y. Fan, P. Lafaye de Micheaux, S. Penev, and D. Salopek. Multivariate nonparametric test of independence. *Journal of Multivariate Analysis*, 153:189–210, 2017.

[10] L. Gan, N. N. Narisetty, and F. Liang. Bayesian regularization for graphical models with unequal shrinkage. *Journal of the American Statistical Association*, (just-accepted), 2018.

[11] C. Genest and B. Rémillard. Test of independence and randomness based on the empirical copula process. *Test*, 13(2):335–369, 2004.

[12] C. Genest, J.-F. Quessy, and B. Rémillard. Asymptotic local efficiency of cramér–von mises tests for multivariate independence. *The Annals of Statistics*, 35(1):166–191, 2007.

[13] A. Gretton, O. Bousquet, A. Smola, and B. Scholkopf. Measuring statistical dependence with hilbert-schmidt norms. In *ALT*, volume 16, pages 63–78. Springer, 2005.

[14] F. Han, S. Chen, and H. Liu. Distribution-free tests of independence in high dimensions. *Biometrika*, 104(4):813–828, 2017.

[15] W. Hoeffding. A non-parametric test of independence. *The annals of mathematical statistics*, pages 546–557, 1948.

[16] Z. Jin and D. S. Matteson. Independent component analysis via energy-based and kernel-based mutual dependence measures. *arXiv preprint arXiv:1805.06639*, 2018.

[17] Z. Jin, X. Yan, and D. S. Matteson. Testing for conditional mean independence with covariates through martingale difference divergence. *arXiv preprint arXiv:1805.06640*, 2018.

[18] Z. Jin, S. Yao, D. S. Matteson, and X. Shao. *EDMeasure: Energy-Based Dependence Measures*, 2018. R package version 1.2.

[19] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[20] I. Kojadinovic and M. Holmes. Tests of independence among continuous random vectors based on cramér–von mises functionals of the empirical copula process. *Journal of Multivariate Analysis*, 100(6):1137–1154, 2009.

[21] D. Leung and M. Drton. Testing independence in high dimensions with sums of rank correlations. *The Annals of Statistics*, 46(1): 280–307, 2018.

[22] Z. Li and T. H. McCormick. An expectation conditional maximization approach for gaussian graphical models. *arXiv preprint arXiv:1709.06970*, 2017.

[23] Z. R. Li, T. H. McCormick, and S. J. Clark. Bayesian joint spike-and-slab graphical lasso. *arXiv preprint arXiv:1805.07051*, 2018.

[24] J. Lu. On finite-population bayesian inferences for 2^K factorial designs with binary outcomes. *arXiv preprint arXiv:1803.04499*, 2018.

[25] J. Lu, P. Ding, and T. Dasgupta. Treatment effects on ordinal outcomes: causal estimands and sharp bounds. *arXiv preprint arXiv:1507.01542*, 2015.

[26] D. S. Matteson and R. S. Tsay. Independent component analysis via distance covariance. *Journal of the American Statistical Association*, 112(518):623–637, 2017.

[27] R. v. Mises. On the asymptotic distribution of differentiable statistical functions. *The annals of mathematical statistics*, 18(3):309–348, 1947.

[28] K. Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

[29] N. Pfister and J. Peters. *dHSIC: Independence Testing via Hilbert Schmidt Independence Criterion*, 2017. R package version 2.0.

[30] N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):5–31, 2018.

[31] X. Shao and J. Zhang. Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, 109(507):1302–1318, 2014.

[32] C. Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

[33] G. J. Székely and M. L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80, 2005.

[34] G. J. Székely and M. L. Rizzo. Brownian distance covariance. *The annals of applied statistics*, 3(4):1236–1265, 2009.

[35] G. J. Székely and M. L. Rizzo. The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117: 193–213, 2013.
[36] G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.
[37] G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.
[38] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.
[39] S. Yao, X. Zhang, and X. Shao. Testing mutual independence in high dimension via distance covariance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):455–480, 2018.
[40] B. Zhang, M. Dundar, and M. Al Hasan. Bayesian non-exhaustive classification a case study: Online name disambiguation using temporal record streams. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1341–1350. ACM, 2016.

Table 1: empirical size ($\alpha = 0.1$) in Example 1 with 1000 repetitions and $d = 2$.

n	$\mathcal{V}_n^2, \mathcal{R}_n, \mathcal{S}_n$	$\mathcal{Q}_n, \mathcal{J}_n, \mathcal{I}_n$	$\mathcal{Q}_n^\star, \mathcal{J}_n^\star$	$\mathcal{I}_n^\star$
25	0.106	0.102	0.108	0.111
30	0.098	0.115	0.086	0.114
35	0.095	0.101	0.084	0.101
50	0.101	0.101	0.111	0.106
70	0.114	0.109	0.090	0.102
100	0.104	0.105	0.118	0.117

Table 2: empirical power ($\alpha = 0.1$) in Example 1 with 1000 repetitions and $d = 2$.

n	$\mathcal{V}_n^2, \mathcal{R}_n, \mathcal{S}_n$	$\mathcal{Q}_n, \mathcal{J}_n, \mathcal{I}_n$	$\mathcal{Q}_n^\star, \mathcal{J}_n^\star$	$\mathcal{I}_n^\star$
25	0.273	0.246	0.160	0.182
50	0.496	0.448	0.259	0.300
100	0.807	0.751	0.442	0.514
150	0.943	0.922	0.604	0.720
200	0.979	-	0.749	0.836
300	1.000	-	0.889	0.954
500	1.000	-	0.978	0.995

Table 3: empirical size ($\alpha = 0.1$) in Example 2 with 1000 repetitions and $d = 2$.

n	$\mathcal{V}_n^2, \mathcal{R}_n, \mathcal{S}_n$	$\mathcal{Q}_n, \mathcal{J}_n, \mathcal{I}_n$	$\mathcal{Q}_n^\star, \mathcal{J}_n^\star$	$\mathcal{I}_n^\star$
25	0.088	0.093	0.091	0.092
30	0.098	0.104	0.108	0.110
35	0.104	0.102	0.104	0.099
50	0.097	0.098	0.093	0.097
70	0.094	0.097	0.089	0.097
100	0.092	0.092	0.114	0.099

Table 4: empirical power ($\alpha = 0.1$) in Example 2 with 1000 repetitions and $d = 2$.

n	$\mathcal{V}_n^2, \mathcal{R}_n, \mathcal{S}_n$	$\mathcal{Q}_n, \mathcal{J}_n, \mathcal{I}_n$	$\mathcal{Q}_n^\star, \mathcal{J}_n^\star$	$\mathcal{I}_n^\star$
25	0.181	0.185	0.141	0.152
50	0.352	0.339	0.200	0.239
100	0.610	0.607	0.372	0.413
150	0.793	0.792	0.474	0.588
200	0.885	-	0.604	0.711
300	0.989	-	0.803	0.892
500	0.999	-	0.953	0.988

Table 5: empirical size ($\alpha = 0.1$) in Example 3 with 1000 repetitions and $d = 3$.

n	BN^h	BN^d	dHSIC	Q_n	$Q_n^\star$	$\mathcal{R}_n$	$\mathcal{S}_n$	$\mathcal{J}_n$	$\mathcal{J}_n^\star$	$\mathcal{I}_n$	$\mathcal{I}_n^\star$
25	0.103	0.106	0.097	0.095	0.103	0.093	0.096	0.101	0.100	0.091	0.101
30	0.114	0.106	0.101	-	0.110	0.110	0.114	0.108	0.118	0.111	0.125
35	0.101	0.095	0.090	-	0.108	0.106	0.102	0.109	0.106	0.104	0.092
50	0.098	0.100	0.106	-	0.083	0.113	0.108	0.110	0.090	0.105	0.085
70	0.114	0.102	0.107	-	0.107	0.104	0.104	0.098	0.101	0.108	0.109
100	0.127	0.125	0.099	-	0.085	0.106	0.108	0.104	0.103	0.109	0.096

Table 6: empirical power ($\alpha = 0.1$) in Example 3 with 1000 repetitions and $d = 3$.

n	BN^h	BN^d	dHSIC	Q_n	$Q_n^\star$	$\mathcal{R}_n$	$\mathcal{S}_n$	$\mathcal{J}_n$	$\mathcal{J}_n^\star$	$\mathcal{I}_n$	$\mathcal{I}_n^\star$
25	0.992	0.998	0.982	0.383	0.220	0.402	0.418	0.360	0.199	0.384	0.228
50	1.000	1.000	1.000	-	0.378	0.707	0.719	0.651	0.338	0.671	0.389
100	1.000	1.000	1.000	-	0.707	0.956	0.961	0.940	0.643	0.946	0.767
150	1.000	1.000	1.000	-	0.873	0.996	0.996	0.993	0.830	0.994	0.921
200	1.000	1.000	1.000	-	0.946	1.000	1.000	-	0.930	-	0.972
300	1.000	1.000	1.000	-	0.997	1.000	1.000	-	0.996	-	0.999
500	1.000	1.000	1.000	-	1.000	1.000	1.000	-	1.000	-	1.000

Table 7: empirical size ($\alpha = 0.1$) in Example 4 with 1000 repetitions and $d = 3$.

n	BN^h	BN^d	dHSIC	Q_n	$Q_n^\star$	$\mathcal{R}_n$	$\mathcal{S}_n$	$\mathcal{J}_n$	$\mathcal{J}_n^\star$	$\mathcal{I}_n$	$\mathcal{I}_n^\star$
25	0.099	0.105	0.102	0.089	0.098	0.096	0.097	0.096	0.099	0.092	0.108
30	0.092	0.089	0.087	-	0.098	0.102	0.100	0.094	0.099	0.095	0.108
35	0.105	0.104	0.087	-	0.116	0.116	0.122	0.123	0.117	0.123	0.113
50	0.098	0.096	0.107	-	0.091	0.112	0.109	0.102	0.097	0.113	0.088
70	0.127	0.127	0.101	-	0.084	0.103	0.105	0.096	0.112	0.102	0.116
100	0.105	0.103	0.110	-	0.112	0.105	0.105	0.109	0.099	0.104	0.107

Table 8: empirical power ($\alpha = 0.1$) in Example 4 with 1000 repetitions and $d = 3$.

n	BN^h	BN^d	dHSIC	Q_n	$Q_n^\star$	$\mathcal{R}_n$	$\mathcal{S}_n$	$\mathcal{J}_n$	$\mathcal{J}_n^\star$	$\mathcal{I}_n$	$\mathcal{I}_n^\star$
25	0.285	0.268	0.267	0.289	0.164	0.294	0.287	0.291	0.154	0.287	0.169
50	0.479	0.479	0.441	-	0.280	0.504	0.510	0.490	0.278	0.501	0.320
100	0.768	0.760	0.745	-	0.521	0.824	0.826	0.807	0.498	0.816	0.579
150	0.919	0.929	0.906	-	0.689	0.942	0.942	0.937	0.679	0.941	0.770
200	0.982	0.987	0.963	-	0.838	0.987	0.986	-	0.826	-	0.905
300	0.999	0.999	0.997	-	0.957	0.999	0.999	-	0.956	-	0.982
500	1.000	1.000	1.000	-	1.000	1.000	1.000	-	1.000	-	1.000

Appendix

Proofs of Theorem 1, 2, 3, 4, 5, 6, and Lemma 1, 2.

Theorem 1

Proof. (i) $0 \le Q(X) < \infty$.

(ii) $Q(X) = 0 \iff X_1, \ldots, X_d$ are mutually independent.

(iii) $Q(X) = \mathrm{E}|X - \widetilde{X}'|_p + \mathrm{E}|X' - \widetilde{X}|_p - \mathrm{E}|X - X'|_p - \mathrm{E}|\widetilde{X} - \widetilde{X}'|_p$.

Since $w_1(t)$ is a positive weight function, $X_1, \ldots, X_d$ are mutually independent if and only if $Q(X) = \int_{\mathbb{R}^p} |\phi_X(t) - \phi_{\widetilde{X}}(t)|^2 w_1(t)\, dt$ is equal to zero.

Table 9: empirical size ($\alpha = 0.1$) in Example 5 with 1000 repetitions and $n = 100$.

d	HL^{τ}	HL^{ρ}	HL_n^{τ}	HL_n^{ρ}	$Q_n^{\star}$	$\mathcal{R}_n$	$\mathcal{S}_n$	$\mathcal{J}_n^{\star}$	$\mathcal{I}_n^{\star}$
5	0.076	0.066	0.113	0.105	0.097	0.091	0.091	0.094	0.104
10	0.077	0.070	0.104	0.097	0.107	0.092	0.094	0.119	0.107
15	0.094	0.087	0.116	0.113	0.109	0.093	0.093	0.108	0.100
20	0.077	0.066	0.089	0.089	0.096	0.099	0.118	0.115	0.101
25	0.074	0.058	0.086	0.091	0.097	0.090	0.082	0.095	0.097
30	0.091	0.082	0.110	0.114	0.109	0.092	0.104	0.105	0.109
50	0.080	0.061	0.088	0.087	0.087	0.091	0.088	0.095	0.087

Table 10: empirical power ($\alpha = 0.1$) in Example 5 with 1000 repetitions and $n = 100$.

d	HL^{τ}	HL^{ρ}	HL_n^{τ}	HL_n^{ρ}	$Q_n^{\star}$	$\mathcal{R}_n$	$\mathcal{S}_n$	$\mathcal{J}_n^{\star}$	$\mathcal{I}_n^{\star}$
5	0.317	0.305	0.410	0.405	0.298	0.545	0.557	0.245	0.318
10	0.426	0.416	0.500	0.510	0.557	0.896	0.915	0.409	0.497
15	0.513	0.481	0.593	0.602	0.822	0.975	0.982	0.538	0.643
20	0.558	0.534	0.625	0.634	0.924	0.996	0.999	0.586	0.647
25	0.593	0.539	0.645	0.634	0.977	0.999	0.999	0.663	0.689
30	0.605	0.556	0.675	0.664	0.980	1.000	1.000	0.711	0.700
50	0.702	0.641	0.742	0.731	0.998	1.000	1.000	0.775	0.717

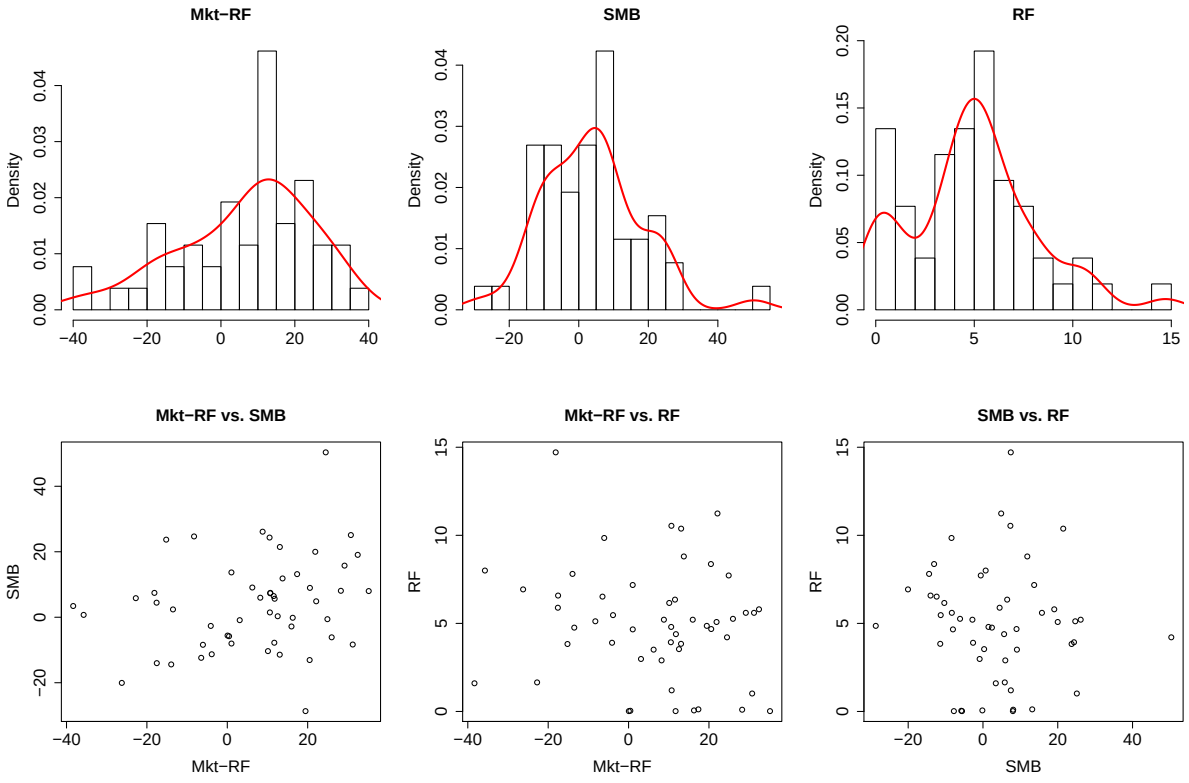

Figure 1: Three annual Fama−French factors between 1964 and 2015: Mkt-RF (excess return on the market), SMB (small minus big) and RF (risk-free return). The correlations are corr(Mkt-RF, SMB) = 0.238, corr(Mkt-RF, RF) = -0.161, and corr(SMB, RF) = -0.0645. Red lines in the histograms are estimated kernel densities.

By the boundedness property of characteristic functions and Fubini's theorem, we have

$$|\phi_X(t) - \phi_{\widetilde{X}}(t)|^2 = \phi_X(t)\overline{\phi_X(t)} + \phi_{\widetilde{X}}(t)\overline{\phi_{\widetilde{X}}(t)} - \phi_X(t)\overline{\phi_{\widetilde{X}}(t)} - \phi_{\widetilde{X}}(t)\overline{\phi_X(t)}$$

$$= [\mathrm{E}^{i\langle t,X\rangle}]\mathrm{E}[e^{-i\langle t,X\rangle}] + \mathrm{E}[e^{i\langle t,\widetilde{X}\rangle}]\mathrm{E}[e^{-i\langle t,\widetilde{X}\rangle}] - \mathrm{E}[e^{i\langle t,X\rangle}]\mathrm{E}[e^{-i\langle t,\widetilde{X}\rangle}] - \mathrm{E}[e^{i\langle t,\widetilde{X}\rangle}]\mathrm{E}[e^{-i\langle t,X\rangle}]$$

$= \mathrm{E}[e^{i\langle t, X-X'\rangle}] + \mathrm{E}[e^{i\langle t, \widetilde{X}-\widetilde{X}'\rangle}] - \mathrm{E}[e^{i\langle t, X-\widetilde{X}'\rangle}] - \mathrm{E}[e^{i\langle t, \widetilde{X}-X'\rangle}]$

$= \mathrm{E}(\cos\langle t, X-X'\rangle) + \mathrm{E}(\cos\langle t, \widetilde{X}-\widetilde{X}'\rangle) + \mathrm{E}(\cos\langle t, X-\widetilde{X}'\rangle) + \mathrm{E}(\cos\langle t, \widetilde{X}-X'\rangle)$

$= \mathrm{E}(1-\cos\langle t, X-\widetilde{X}'\rangle) + \mathrm{E}(1-\cos\langle t, \widetilde{X}-X'\rangle)$

$\quad - \mathrm{E}(1-\cos\langle t, X-X'\rangle) - \mathrm{E}(1-\cos\langle t, \widetilde{X}-\widetilde{X}'\rangle).$

Since $\mathrm{E}|X|_p < \infty$ implies $\mathrm{E}|\widetilde{X}|_p < \infty$, we have $\mathrm{E}(|X|_p + |\widetilde{X}|_p) < \infty$. Then the triangle inequality implies $\mathrm{E}|X-X'|_p, \mathrm{E}|\widetilde{X}-\widetilde{X}'|_p, \mathrm{E}|X-\widetilde{X}'|_p, \mathrm{E}|\widetilde{X}-X'|_p < \infty$. Therefore, by Fubini's theorem and Lemma 1, it follows that

$Q(X) = \int |\phi_X(t) - \phi_{\widetilde{X}}(t)|^2 \, w_1(t)\, dt$

$= \int \mathrm{E}(1-\cos\langle t, X-\widetilde{X}'\rangle)\, w_1(t)\, dt + \int \mathrm{E}(1-\cos\langle t, \widetilde{X}-X'\rangle)\, w_1(t)\, dt$

$\quad - \int \mathrm{E}(1-\cos\langle t, X-X'\rangle)\, w_1(t)\, dt - \int \mathrm{E}(1-\cos\langle t, \widetilde{X}-\widetilde{X}'\rangle)\, w_1(t)\, dt$

$= \mathrm{E}|X-\widetilde{X}'|_p + \mathrm{E}|\widetilde{X}-X'|_p - \mathrm{E}|X-X'|_p - \mathrm{E}|\widetilde{X}-\widetilde{X}'|_p < \infty.$

Finally, $Q(X) \geq 0$ since the integrand $|\phi_X(t) - \phi_{\widetilde{X}}(t)|^2$ is non-negative. □

Lemma 1

Proof. After a simple calculation, we have

$|\phi_X^n(t) - \phi_{\widetilde{X}}^n(t)|^2 = \phi_X^n(t)\overline{\phi_X^n(t)} - \phi_X^n(t)\overline{\phi_{\widetilde{X}}^n(t)} - \phi_{\widetilde{X}}^n(t)\overline{\phi_X^n(t)} + \phi_{\widetilde{X}}^n(t)\overline{\phi_{\widetilde{X}}^n(t)}$

$= \frac{1}{n^2}\sum_{k,\ell=1}^n \cos\langle t, X^k - X^\ell\rangle - \frac{2}{n^{d+1}}\sum_{k,\ell_1,\ldots,\ell_d=1}^n \cos\langle t, X^k - (X_1^{\ell_1}, \ldots, X_d^{\ell_d})\rangle$

$\quad + \frac{1}{n^{2d}}\sum_{k_1,\ldots,k_d,\ell_1,\ldots,\ell_d=1}^n \cos\langle t, (X_1^{k_1}, \ldots, X_d^{k_d}) - (X_1^{\ell_1}, \ldots, X_d^{\ell_d})\rangle + V$

$= -\frac{1}{n^2}\sum_{k,\ell=1}^n [1-\cos\langle t, X^k - X^\ell\rangle] + \frac{2}{n^{d+1}}\sum_{k,\ell_1,\ldots,\ell_d=1}^n [1-\cos\langle t, X^k - (X_1^{\ell_1}, \ldots, X_d^{\ell_d})\rangle]$

$\quad - \frac{1}{n^{2d}}\sum_{k_1,\ldots,k_d,\ell_1,\ldots,\ell_d=1}^n [1-\cos\langle t, (X_1^{k_1}, \ldots, X_d^{k_d}) - (X_1^{\ell_1}, \ldots, X_d^{\ell_d})\rangle] + V,$

where V is imaginary and thus 0 as the $|\phi_X^n(t) - \phi_{\widetilde{X}}^n(t)|^2$ is real.

By Lemma 1 in Székely and Rizzo [33]

$Q_n(\mathbf{X}) = \|\phi_X^n(t) - \phi_{\widetilde{X}}^n(t)\|_{w_1}^2$

$= -\frac{1}{n^2}\sum_{k,\ell=1}^n |X^k - X^\ell|_p + \frac{2}{n^{d+1}}\sum_{k,\ell_1,\ldots,\ell_d=1}^n |X^k - (X_1^{\ell_1}, \ldots, X_d^{\ell_d})|_p$

$\quad - \frac{1}{n^{2d}}\sum_{k_1,\ldots,k_d,\ell_1,\ldots,\ell_d=1}^n |(X_1^{k_1}, \ldots, X_d^{k_d}) - (X_1^{\ell_1}, \ldots, X_d^{\ell_d})|_p.$ □

Lemma 2

Proof. After a simple calculation, we have

$|\phi_X^n(t) - \phi_{\widetilde{X}}^{n\star}(t)|^2 = \phi_X^n(t)\overline{\phi_X^n(t)} - \phi_X^n(t)\overline{\phi_{\widetilde{X}}^{n\star}(t)} - \phi_{\widetilde{X}}^{n\star}(t)\overline{\phi_X^n(t)} + \phi_{\widetilde{X}}^{n\star}(t)\overline{\phi_{\widetilde{X}}^{n\star}(t)}$

$= \frac{1}{n^2}\sum_{k,\ell=1}^n \cos\langle t, X^k - X^\ell\rangle - \frac{2}{n^2}\sum_{k,\ell=1}^n \cos\langle t, X^k - (X_1^\ell, \ldots, X_d^{\ell+d-1})\rangle$

$\quad + \frac{1}{n^2}\sum_{k,\ell=1}^n \cos\langle t, (X_1^k, \ldots, X_d^{k+d-1}) - (X_1^\ell, \ldots, X_d^{\ell+d-1})\rangle + V^\star$

$= -\frac{1}{n^2}\sum_{k,\ell=1}^n [1-\cos\langle t, X^k - X^\ell\rangle] + \frac{2}{n^2}\sum_{k,\ell=1}^n [1-\cos\langle t, X^k - (X_1^\ell, \ldots, X_d^{\ell+d-1})\rangle]$

$\quad - \frac{1}{n^2}\sum_{k,\ell=1}^n [1-\cos\langle t, (X_1^k, \ldots, X_d^{k+d-1}) - (X_1^\ell, \ldots, X_d^{\ell+d-1})\rangle] + V^\star,$

where $V^\star$ is imaginary and thus 0 as the $|\phi_X^n(t) - \phi_{\widetilde{X}}^{n\star}(t)|^2$ is real.

By Lemma Lemma 1 in Székely and Rizzo [33]

$$
\begin{aligned}
&Q_n^\star(\mathbf{X}) = \|\phi_X^n(t) - \phi_{\widetilde{X}}^{n\star}(t)\|_{w_1}^2 \\
&= -\tfrac{1}{n^2}\sum_{k,\ell=1}^n |X^k - X^\ell|_p + \tfrac{2}{n^2}\sum_{k,\ell=1}^n |X^k - (X_1^\ell,\ldots,X_d^{\ell+d-1})|_p \\
&\quad - \tfrac{1}{n^2}\sum_{k,\ell=1}^n |(X_1^k,\ldots,X_d^{k+d-1}) - (X_1^\ell,\ldots,X_d^{\ell+d-1})|_p. \qquad \square
\end{aligned}
$$

Theorem 2

Proof. We define

$$
Q_n = \|\phi_X^n(t) - \phi_{\widetilde{X}}^n(t)\|_{w_1}^2 \triangleq \|\xi_n(t)\|_{w_1}^2 \quad \text{and} \quad Q_n^\star = \|\phi_X^n(t) - \phi_{\widetilde{X}}^{n\star}(t)\|_{w_1}^2 \triangleq \|\xi_n^\star(t)\|_{w_1}^2.
$$

For $\forall 0 < \delta < 1$, define the region

$$
D(\delta) = \{t = (t_1,\ldots,t_d) : \delta \le |t|_p^2 = \sum_{j=1}^d |t_j|_{p_j}^2 \le 2/\delta\}, \tag{3}
$$

and random variables

$$
Q_{n,\delta} = \int_{D(\delta)} |\xi_n(t)|^2\, dw_1 \quad \text{and} \quad Q_{n,\delta}^\star = \int_{D(\delta)} |\xi_n^\star(t)|^2\, dw_1.
$$

For any fixed δ, the weight function $w_1(t)$ is bounded on $D(\delta)$. Hence $Q_{n,\delta}$ is a combination of V-statistics of bounded random variables. Similar to Theorem 2 of Székely et al. [37], it follows by the strong law of large numbers (SLLN) for V-statistics [27] that almost surely

$$
\lim_{n\to\infty} Q_{n,\delta} = \lim_{n\to\infty} Q_{n,\delta}^\star = Q_{\cdot,\delta} = \int_{D(\delta)} |\phi_X(t) - \phi_{\widetilde{X}}(t)|^2\, dw_1.
$$

Clearly $Q_{\cdot,\delta} \to Q$ as $\delta \to 0$. Hence, $Q_{n,\delta} \to Q$ a.s. and $Q_{n,\delta}^\star \to Q$ a.s. as $\delta \to 0$, $n \to \infty$. In order to show $Q_n \to Q$ a.s. and $Q_n^\star \to Q$ a.s. as $n \to \infty$, it remains to prove that almost surely

$$
\limsup_{\delta\to 0}\limsup_{n\to\infty} |Q_{n,\delta} - Q_n| = \limsup_{\delta\to 0}\limsup_{n\to\infty} |Q_{n,\delta}^\star - Q_n^\star| = 0.
$$

We define a mixture of $\widetilde{X}$ and X as $Y_{-c} = (\widetilde{X}_1,\ldots,\widetilde{X}_{c-1}, X_{c^+})$, $c = 1,\ldots,d-1$.

By the Cauchy−Bunyakovsky inequality

$$
\begin{aligned}
&|\xi_n(t)|^2 = |\phi_X^n(t) - \textstyle\prod_{j=1}^d \phi_{X_j}^n(t_j)|^2 \\
&= |\phi_X^n(t) - \textstyle\prod_{j=1}^d \phi_{X_j}^n(t_j) - \sum_{c=1}^{d-2}(\prod_{j=1}^c \phi_{X_j}^n(t_j)\phi_{X_{c^+}}^n(t_{c^+})) + \sum_{c=1}^{d-2}(\prod_{j=1}^c \phi_{X_j}^n(t_j)\phi_{X_{c^+}}^n(t_{c^+}))|^2 \\
&\le [|\phi_X^n(t) - \phi_{X_1}^n(t_1)\phi_{X_{1^+}}^n(t_{1^+})| \\
&\quad + \textstyle\sum_{c=1}^{d-2} |(\prod_{j=1}^c \phi_{X_j}^n(t_j)\phi_{X_{c^+}}^n(t_{c^+})) - (\prod_{j=1}^c \phi_{X_j}^n(t_j)\phi_{X_{c+1}}^n(t_{c+1})\phi_{X_{(c+1)^+}}^n(t_{(c+1)^+}))|]^2 \\
&= [\textstyle\sum_{c=1}^{d-1} |\phi_{(X_c,Y_{-c})}^n(t_c,t_{-c}) - \phi_{X_c}^n(t_c)\phi_{Y_{-c}}^n(t_{-c})|]^2
\end{aligned}
$$

$$\leq (d-1)\sum_{c=1}^{d-1}|\phi^n_{(X_c,Y_{-c})}(t_c,t_{-c})-\phi^n_{X_c}(t_c)\phi^n_{Y_{-c}}(t_{-c})|^2,$$

and

$$|\xi^\star_n(t)|^2=|\tfrac{1}{n}\sum_{k=1}^n e^{i\langle t,X^k\rangle}-\tfrac{1}{n}\sum_{k=1}^n e^{i\sum_{j=1}^d\langle t_j,X_j^{k+j-1}\rangle}|^2$$
$$=|\tfrac{1}{n}\sum_{k=1}^n(e^{i\langle t,X^k\rangle}-\sum_{c=2}^{d-1}e^{i\langle t,(X_1^k,\ldots,X_c^{k+c-1},X^k_{c^+})\rangle}+\sum_{c=2}^{d-1}e^{i\langle t,(X_1^k,\ldots,X_c^{k+c-1},X^k_{c^+})\rangle}-e^{i\sum_{j=1}^d\langle t_j,X_j^{k+j-1}\rangle})|^2$$
$$=|\tfrac{1}{n}\sum_{k=1}^n\sum_{c=1}^{d-1}(e^{i\langle t,(X_1^k,\ldots,X_c^{k+c-1},X^k_{c^+})\rangle}-e^{i\langle t,(X_1^k,\ldots,X_{c+1}^{k+c},X^k_{(c+1)^+})\rangle})|^2$$
$$\leq(d-1)\sum_{c=1}^{d-1}|\tfrac{1}{n}\sum_{k=1}^n e^{i\langle t_{-(c+1)},(X_1^k,\ldots,X_c^{k+c-1},X^k_{(c+1)^+})\rangle}(e^{i\langle t_{c+1},X^k_{c+1}\rangle}-e^{i\langle t_{c+1},X^{k+c}_{c+1}\rangle})|^2$$
$$\leq(d-1)\sum_{c=1}^{d-1}(\tfrac{1}{n}\sum_{k=1}^n|e^{i\langle t_{-(c+1)},(X_1^k,\ldots,X_c^{k+c-1},X^k_{(c+1)^+})\rangle}|^2\tfrac{1}{n}\sum_{k=1}^n|e^{i\langle t_{c+1},X^k_{c+1}\rangle}-e^{i\langle t_{c+1},X^{k+c}_{c+1}\rangle}|^2)$$
$$=(d-1)\sum_{c=1}^{d-1}(\tfrac{1}{n}\sum_{k=1}^n|e^{i\langle t_{c+1},X^k_{c+1}\rangle}-e^{i\langle t_{c+1},X^{k+c}_{c+1}\rangle}|^2)$$
$$\leq(d-1)\sum_{c=2}^{d}\tfrac{2}{n}\sum_{k=1}^n(|e^{i\langle t_c,X_c^k\rangle}-\phi_{X_c}(t_c)|^2+|\phi_{X_c}(t_c)-e^{i\langle t_c,X_c^{k+c-1}\rangle}|^2)$$
$$=4(d-1)\sum_{c=2}^{d}\tfrac{1}{n}\sum_{k=1}^n|e^{i\langle t_c,X_c^k\rangle}-\phi_{X_c}(t_c)|^2.$$

By the inequality $sa+(1-s)b\geq a^sb^{1-s}$, $0<s<1$, $a,b>0$, we have

$$|t|_p^{1+p}=(|t_c|_{p_c}^2+|t_{-c}|_{p_{-c}}^2)^{\frac{1+p}{2}}\geq(\tfrac{1+p_c}{2+p}|t_c|_{p_c}^2+\tfrac{1+p_{-c}}{2+p}|t_{-c}|_{p_{-c}}^2)^{\frac{1+p}{2}}\geq(|t_c|_{p_c}^{\frac{2(1+p_c)}{2+p}}|t_{-c}|_{p_{-c}}^{\frac{2(1+p_{-c})}{2+p}})^{\frac{1+p}{2}}$$
$$=|t_c|_{p_c}^{\frac{1+p_{-c}}{2+p}+p_c}|t_{-c}|_{p_{-c}}^{\frac{1+p_c}{2+p}+p_{-c}}\triangleq|t_c|_{p_c}^{m_c+p_c}|t_{-c}|_{p_{-c}}^{m_{-c}+p_{-c}},$$

where $p_{-c}=\sum_{j\neq c}p_j=p-p_c$, $0<m_c<1$, $0<m_{-c}<1$ and consequently

$$w_1(t)=\tfrac{1}{K(p,1)|t|_p^{1+p}}\leq\tfrac{K(p_c,m_c)K(p_{-c},m_{-c})}{K(p,1)}\tfrac{1}{K(p_c,m_c)|t_c|_{p_c}^{m_c+p_c}}\tfrac{1}{K(p_{-c},m_{-c})|t_{-c}|_{p_{-c}}^{m_{-c}+p_{-c}}}$$
$$\triangleq C(p,p_c,p_{-c})\tfrac{1}{K(p_c,m_c)|t_c|_{p_c}^{m_c+p_c}}\tfrac{1}{K(p_{-c},m_{-c})|t_{-c}|_{p_{-c}}^{m_{-c}+p_{-c}}},$$

where $C(p,p_c,p_{-c})$ is a constant depending only on p,p_c,p_{-c}.

By the fact $\{\mathbb{R}^p\backslash D(\delta)\}\subset\{|t_c|_{p_c}^2,|t_{-c}|_{p_{-c}}^2<\delta\}\cup\{|t_c|_{p_c}^2>1/\delta\}\cup\{|t_{-c}|_{p_{-c}}^2>1/\delta\}$ and similar steps in Theorem 2 of Székely et al. [37], almost surely

$$\limsup_{\delta\to0}\limsup_{n\to\infty}|Q_{n,\delta}-Q_n|=\limsup_{\delta\to0}\limsup_{n\to\infty}\int_{\mathbb{R}^p\backslash D(\delta)}|\xi_n(t)|^2\,dw_1$$
$$\leq(d-1)\sum_{c=1}^{d-1}\limsup_{\delta\to0}\limsup_{n\to\infty}\int_{\mathbb{R}^p\backslash D(\delta)}|\phi^n_{(X_c,Y_{-c})}(t_c,t_{-c})-\phi^n_{X_c}(t_c)\phi^n_{Y_{-c}}(t_{-c})|^2\,dw_1$$
$$\leq C(p,p_c,p_{-c})(d-1)\sum_{c=1}^{d-1}\limsup_{\delta\to0}\limsup_{n\to\infty}\int_{\mathbb{R}^p\backslash D(\delta)}|\phi^n_{(X_c,Y_{-c})}(t_c,t_{-c})$$
$$-\phi^n_{X_c}(t_c)\phi^n_{Y_{-c}}(t_{-c})|^2\tfrac{1}{K(p_c,m_c)|t_c|_{p_c}^{m_c+p_c}}\tfrac{1}{K(p_{-c},m_{-c})|t_{-c}|_{p_{-c}}^{m_{-c}+p_{-c}}}\,dt_c\,dt_{-c}$$
$$=0,$$

and

$$\limsup_{\delta\to0}\limsup_{n\to\infty}|Q^\star_{n,\delta}-Q^\star_n|=\limsup_{\delta\to0}\limsup_{n\to\infty}\int_{\mathbb{R}^p\backslash D(\delta)}|\xi^\star_n(t)|^2\,dw_1$$
$$\leq4(d-1)\sum_{c=2}^{d}\limsup_{\delta\to0}\limsup_{n\to\infty}\tfrac{1}{n}\sum_{k=1}^n\int_{\mathbb{R}^p\backslash D(\delta)}|e^{i\langle t_c,X_c^k\rangle}-\phi_{X_c}(t_c)|^2\,dw_1$$
$$\leq C(p,p_c,p_{-c})4(d-1)\sum_{c=2}^{d}\limsup_{\delta\to0}\limsup_{n\to\infty}\tfrac{1}{n}\sum_{k=1}^n\int_{\mathbb{R}^p\backslash D(\delta)}|e^{i\langle t_c,X_c^k\rangle}-\phi_{X_c}(t_c)|^2$$
$$\tfrac{1}{K(p_c,m_c)|t_c|_{p_c}^{m_c+p_c}}\tfrac{1}{K(p_{-c},m_{-c})|t_{-c}|_{p_{-c}}^{m_{-c}+p_{-c}}}\,dt_c\,dt_{-c}$$
$$=0.$$

Therefore, almost surely

$$\limsup_{\delta\to0}\limsup_{n\to\infty}|Q_{n,\delta}-Q_n|=\limsup_{\delta\to0}\limsup_{n\to\infty}|Q^\star_{n,\delta}-Q^\star_n|=0.\qquad\square$$

Theorem 3

Proof. (i) Under H_0:

Let $\zeta(t)$ denote a complex-valued Gaussian processe with mean zero and covariance functions

$$\begin{aligned} R(t,t^0) &= \prod_{j=1}^{d}\phi_{X_j}(t_j-t_j^0)+(d-1)\prod_{j=1}^{d}\phi_{X_j}(t_j)\overline{\phi_{X_j}(t_j^0)} \\ &\quad -\sum_{j=1}^{d}\phi_{X_j}(t_j-t_j^0)\prod_{j'\neq j}\phi_{X_{j'}}(t_{j'})\overline{\phi_{X_{j'}}(t_{j'}^0)}. \end{aligned}$$

We define

$$n\mathcal{Q}_n = n\|\phi_X^n(t)-\phi_{\tilde{X}}^n(t)\|_{w_1}^2 \triangleq \|\zeta_n(t)\|_{w_1}^2.$$

After a simple calculation, we have

$\mathrm{E}[\zeta_n(t)] = \mathrm{E}[\zeta_n^\star(t)] = 0$,

$\mathrm{E}[\zeta_n(t)\overline{\zeta_n(t^0)}]$

$= (1-\frac{1}{n^{d-1}})\prod_{j=1}^{d}\phi_{X_j}(t_j-t_j^0)+(n-1-\frac{(n-1)^d}{n^{d-1}})\prod_{j=1}^{d}\phi_{X_j}(t_j)\overline{\phi_{X_j}(t_j^0)}$

$\quad -\frac{(n-1)^{d-1}}{n^{d-1}}[\sum_{j=1}^{d}\phi_{X_j}(t_j-t_j^0)\prod_{j'\neq j}\phi_{X_j}(t_j)\overline{\phi_{X_j}(t_j^0)}]+o_n(1)$

$\to R(t,t^0)$ as $n\to\infty$.

In particular, $\mathrm{E}|\zeta_n(t)|^2 \to R(t,t) \le d$ as $n\to\infty$. Thus, $\mathrm{E}|\zeta_n(t)|^2 \le d+1$ for enough large n.

For $\forall 0<\delta<1$, define the region $D(\delta)$ as (3). Given $\forall\epsilon>0$, we choose a partition $\{D^\ell(\delta)\}_{\ell=1}^N$ of $D(\delta)$ into $N(\epsilon)$ measurable sets with diameter at most ϵ, and suppress the notation of $D(\delta), D^\ell(\delta)$ as D, D^ℓ. Then we define two sequences of random variables for any fixed $t^\ell\in D^\ell, \ell=1,\dots,N$

$$Q_n(\delta) = \sum_{\ell=1}^{N}\int_{D^\ell}|\zeta_n(t^\ell)|^2\,dw_1.$$

For any fixed $M>0$, let $\beta(\epsilon)=\sup_{t,t^0}E||\zeta_n(t)|^2-|\zeta_n(t^0)|^2|$ where the supremum is taken over all $t=(t_1,\dots,t_d)$ and $t^0=(t_1^0,\dots,t_d^0)$ s.t. $\max\{|t|_p^2,|t^0|_p^2\}\le M$ and $|t-t^0|_p^2=\sum_{j=1}^d|t_j-t_j^0|_p^2\le\epsilon^2$. By the continuous mapping theorem and $\zeta_n(t)\to\zeta_n(t^0)$ as $\epsilon\to0$, we have $|\zeta_n(t)|^2\to|\zeta_n(t^0)|^2$ as $\epsilon\to0$. By the dominated convergence theorem and $\mathrm{E}|\zeta_n(t)|^2\le d+1$ for enough large n, we have $E||\zeta_n(t)|^2-|\zeta_n(t^0)|^2|\to0$ as $\epsilon\to0$, which leads to $\beta(\epsilon)\to0$ as $\epsilon\to0$.

As a result

$\mathrm{E}|\int_D|\zeta_n(t)|^2\,dw_1-Q_n(\delta)| = \mathrm{E}|\sum_{\ell=1}^N\int_{D^\ell}(|\zeta_n(t)|^2-|\zeta_n(t^\ell)|^2)\,dw_1|$

$\le\sum_{\ell=1}^N\int_{D^\ell}\mathrm{E}||\zeta_n(t)|^2-|\zeta_n(t^\ell)|^2|\,dw_1\le\beta(\epsilon)\int_D 1\,dw_1$

$\to0$ as $\epsilon\to0$.

By similar steps in Theorem 2, we have

$\mathrm{E}|\int_D|\zeta_n(t)|^2\,dw_1-\|\zeta_n\|_{w_1}^2|\to0$ as $\delta\to0$ and $\mathrm{E}|\int_D|\zeta_n^\star(t)|^2\,dw_1-\|\zeta_n^\star\|_{w_1}^2|\to0$ as $\delta\to0$.

Therefore

$\mathrm{E}|Q_n(\delta) - \|\zeta_n\|^2_{w_1}| \to 0$ as $\epsilon, \delta \to 0$ and $\mathrm{E}|Q^\star_n(\delta) - \|\zeta^\star_n\|^2_{w_1}| \to 0$ as $\epsilon, \delta \to 0$.

On the other hand, define two random variables for any fixed $t^\ell \in D^\ell$, $\ell = 1, \ldots, N$

$$Q(\delta) = \sum_{\ell=1}^{N} \int_{D^\ell} |\zeta(t^\ell)|^2 \, dw_1.$$

Similarly, we have

$\mathrm{E}|Q(\delta) - \|\zeta\|^2_{w_1}| \to 0$ as $\epsilon, \delta \to 0$.

By the multivariate central limit theorem, delta method and continuous mapping theorem, we have

$Q_n(\delta) = \sum_{\ell=1}^{N} \int_{D^\ell} |\zeta_n(t^\ell)|^2 \, dw_1 \to_{\mathcal{D}} \sum_{\ell=1}^{N} \int_{D^\ell} |\zeta(t^\ell)|^2 \, dw_1 = Q(\delta)$ as $n \to \infty$.

Therefore

$\|\zeta_n\|^2_{w_1} \to_{\mathcal{D}} \|\zeta\|^2_{w_1}$ as $\epsilon, \delta \to 0$, $n \to \infty$,

since $\{Q_n(\delta)\}$ have the following properties

(a) $Q_n(\delta)$ converges in distribution to $Q(\delta)$ as $n \to \infty$.

(b) $\mathrm{E}|Q_n(\delta) - \|\zeta_n\|^2_{w_1}| \to 0$ as $\epsilon, \delta \to 0$.

(c) $\mathrm{E}|Q(\delta) - \|\zeta\|^2_{w_1}| \to 0$ as $\epsilon, \delta \to 0$.

Analogous to $\zeta(t), \zeta_n(t), \beta(\epsilon), Q(\delta)$, $Q_n(\delta)$ for $\mathcal{Q}_n$, we can define $\zeta^\star(t), \zeta^\star_n(t), \beta^\star(\epsilon), Q^\star(\delta)$, $Q^\star_n(\delta)$ for $\mathcal{Q}^\star_n$, and prove that $\|\zeta^\star_n\|^2_{w_1} \to_{\mathcal{D}} \|\zeta^\star\|^2_{w_1}$ as $\epsilon, \delta \to 0$, $n \to \infty$ through the same derivations. The only differences are $\mathrm{E}[\zeta^\star_n(t)\overline{\zeta^\star_n(t^0)}] = 2R(t, t^0)$ and $\mathrm{E}|\zeta^\star_n(t)|^2 = 2R(t, t) \le 2d + 1$ for enough large n.

(ii) Under H_A:

By Theorem 1 and 2, we have

$\mathcal{Q}_n \to \mathcal{Q} > 0$ a.s. as $n \to \infty$.

Therefore

$n\mathcal{Q}_n \to \infty$ a.s. as $n \to \infty$.

Similarly, we can prove that $n\mathcal{Q}^\star_n \to \infty$ a.s. as $n \to \infty$ through the same derivations. □

Theorem 4

Proof. (i) $0 \le \mathcal{R}(X) < \infty$.

(ii) $0 \le \mathcal{S}(X) < \infty$.

(iii) $\mathcal{R}(X) = \sum_{c=1}^{d-1} \mathcal{V}^2(X_c, X_{c^+}) = 0 \iff X_1, \ldots, X_d$ are mutually independent.

(iv) $\mathcal{S}(X) = \sum_{c=1}^{d} \mathcal{V}^2(X_c, X_{-c}) = 0 \iff X_1, \ldots, X_d$ are mutually independent.

Since $\mathrm{E}|X|_p < \infty$, we have $0 \le \mathcal{V}^2(X_c, X_{c^+}) < \infty$, $c = 1, \ldots, d-1$. Thus, $0 \le \mathcal{R}(X) = \sum_{c=1}^{d-1} \mathcal{V}^2(X_c, X_{c^+}) < \infty$.

Similarly, we have $0 \le \mathcal{S}(X) = \sum_{c=1}^{d} \mathcal{V}^2(X_c, X_{-c}) < \infty$.

"$\Longleftarrow$"

If $X_1, \ldots, X_d$ are mutually independent, then X_c and X_{c^+} are independent, $\forall c = 1, \ldots, d-1$.

By Theorem 3 of Székely et al. [37], $\mathcal{V}^2(X_c, X_{c^+}) = 0$, $\forall c = 1, \ldots, d-1$.

As a result, $\mathcal{R}(X) = 0$.

Similarly, we can prove that $\mathcal{S}(X) = 0$, since X_c and X_{-c} are independent, $\forall c = 1, \ldots, d$.

"$\Longrightarrow$"

If $\mathcal{R}(X) = 0$, then $\mathcal{V}^2(X_c, X_{c^+}) = 0$, $\forall c = 1, \ldots, d-1$.

By Theorem 3 of Székely et al. [37], X_c and X_{c^+} are independent, $\forall c = 1, \ldots, d-1$. Thus, For all $t \in \mathbb{R}^p$, we have

$$\phi_{(X_j, X_{j^+})}(t_j, t_{j^+}) - \phi_{X_j}(t_j)\phi_{X_{j^+}}(t_{j^+}) = 0,$$

where ϕ_{X_j} and $\phi_{X_{j+}}$ denote the marginal and $\phi_{(X_j, X_{j^+})}$ denotes the joint characteristic function of X_j and X_{j^+} respectively, $j = 1, \ldots, d$.

For all $t \in \mathbb{R}^p$, we have

$|\phi_X(t) - \prod_{j=1}^{d} \phi_{X_j}(t_j)|$

$= |\phi_X(t) - \prod_{j=1}^{d} \phi_{X_j}(t_j) - \sum_{c=1}^{d-2}(\prod_{j=1}^{c} \phi_{X_j}(t_j)\phi_{X_{c^+}}(t_{c^+})) + \sum_{c=1}^{d-2}(\prod_{j=1}^{c} \phi_{X_j}(t_j)\phi_{X_{c^+}}(t_{c^+}))|$

$\leq |\phi_X(t) - \phi_{X_1}(t_1)\phi_{X_{1^+}}(t_{1^+})|$

$\quad + \sum_{c=1}^{d-2} |\prod_{j=1}^{c} \phi_{X_j}(t_j)\phi_{X_{c^+}}(t_{c^+}) - \prod_{j=1}^{c} \phi_{X_j}(t_j)\phi_{X_{c+1}}(t_{c+1})\phi_{X_{(c+1)^+}}(t_{(c+1)^+})|$

$\leq |\phi_X(t) - \phi_{X_1}(t_1)\phi_{X_{1^+}}(t_{1^+})|$

$\quad + \sum_{c=1}^{d-2} |\prod_{j=1}^{c} \phi_{X_j}(t_j)||\phi_{X_{c^+}}(t_{c^+}) - \phi_{X_{c+1}}(t_{c+1})\phi_{X_{(c+1)^+}}(t_{(c+1)^+})|$

$\leq |\phi_X(t) - \phi_{X_1}(t_1)\phi_{X_{1^+}}(t_{1^+})| + \sum_{c=1}^{d-2} |\phi_{X_{c^+}}(t_{c^+}) - \phi_{X_{c+1}}(t_{c+1})\phi_{X_{(c+1)^+}}(t_{(c+1)^+})|$

$= \sum_{c=1}^{d-1} |\phi_{(X_c, X_{c^+})}(t_c, t_{c^+}) - \phi_{X_c}(t_c)\phi_{X_{c^+}}(t_{c^+})|$

$= 0.$

Therefore, for all $t \in \mathbb{R}^p$, we have $|\phi_X(t) - \prod_{j=1}^{d} \phi_{X_j}(t_j)| = 0$, which implies that $X_1, \ldots, X_d$ are mutually independent.

Similarly, we can prove that $\mathcal{S}(X) = 0$ implies that $X_1, \ldots, X_d$ are mutually independent, since X_c and X_{-c} are independent implies that X_c and X_{c^+} are independent. □

Theorem 5

Proof. By Theorem 2 of Székely et al. [37]

$\lim_{n\to\infty} \mathcal{V}_n^2(\mathbf{X_c}, \mathbf{X_{c^+}}) = \mathcal{V}^2(X_c, X_{c^+})$, $c = 1, \ldots, d-1$,

$\lim_{n\to\infty} \mathcal{V}_n^2(\mathbf{X_c}, \mathbf{X_{-c}}) = \mathcal{V}^2(X_c, X_{-c})$, $c = 1, \ldots, d$.

Therefore, the limit of sum converges to the sum of limit as

$\mathcal{R}_n(\mathbf{X}) \xrightarrow[n\to\infty]{a.s.} \mathcal{R}(X)$ and $\mathcal{S}_n(\mathbf{X}) \xrightarrow[n\to\infty]{a.s.} \mathcal{S}(X)$. □

Theorem 6

Proof. (i) Under H_0:

We define

$$n\mathcal{R}_n(\mathbf{X}) = n\sum_{c=1}^{d-1}\mathcal{V}_n^2(\mathbf{X}_c, \mathbf{X}_{c^+}) \triangleq \sum_{c=1}^{d-1}\|\zeta_n^c(t_{(c-1)^+})\|_{w_0}^2,$$

which is the sum corresponding to the pairs $\{X_{d-1}, X_d\}, \{X_{d-2}, (X_{d-1}, X_d)\}, \{X_{d-3}, (X_{d-2}, X_{d-1}, X_d)\}, \ldots, \{X_1, (X_2, \ldots, X_d)\}$. Any two of them can be reorganized as $\{X_1, X_2\}$ and $\{X_4, (X_1, X_2, X_3)\}$ where X_3 could be empty. Without loss of generality, next we will show $\phi^n_{(X_1,X_2)}(t_1, t_2) - \phi^n_{X_1}(t_1)\phi^n_{X_2}(t_2)$ and $\phi^n_{(X_1,X_2,X_3,X_4)}(s_1, s_2) - \phi^n_{(X_1,X_2,X_3)}(s_1)\phi^n_{X_4}(s_2)$ are uncorrelated. Then it follows that $\zeta_n^c(t_{(c-1)^+})$, $c = 1, \ldots, d-1$ are uncorrelated.

After a simple calculation, we have

$\mathrm{E}[\phi^n_{(X_1,X_2)}(t_1, t_2) - \phi^n_{X_1}(t_1)\phi^n_{X_2}(t_2)] = \mathrm{E}[\phi^n_{(X_1,X_2,X_3,X_4)}(s_1, s_2) - \phi^n_{(X_1,X_2,X_3)}(s_1)\phi^n_{X_4}(s_2)] = 0,$

$\mathrm{E}[\phi^n_{(X_1,X_2)}(t_1, t_2) - \phi^n_{X_1}(t_1)\phi^n_{X_2}(t_2)][\overline{\phi^n_{(X_1,X_2,X_3,X_4)}(s_1, s_2) - \phi^n_{(X_1,X_2,X_3)}(s_1)\phi^n_{X_4}(s_2)}] = 0.$

As a result

$\mathrm{Cov}(\phi^n_{(X_1,X_2)}(t_1, t_2) - \phi^n_{X_1}(t_1)\phi^n_{X_2}(t_2), \overline{\phi^n_{(X_1,X_2,X_3,X_4)}(s_1, s_2) - \phi^n_{(X_1,X_2,X_3)}(s_1)\phi^n_{X_4}(s_2)}) = 0.$

Let $p_{(c-1)^+} = \sum_{j=c}^d p_j$.

For $\forall\delta > 0$, define the region $D_c(\delta) = \{t_{(c-1)^+} = (t_c, t_{c^+}) = (t_c, \ldots, t_d) : \delta \le |t_{(c-1)^+}|^2_{p_{(c-1)^+}} = \sum_{j=c}^d |t_j|^2_{p_j} \le 2/\delta\}$. Given $\forall\epsilon > 0$, we choose a partition $\{D_c^\ell\}_{\ell=1}^{N_c}$ of $D_c(\delta)$ into $N_c(\epsilon)$ measurable sets with diameter at most ϵ, and define a sequence of random variables for any fixed $t^\ell_{(c-1)^+} \in D_c^\ell$, $\ell = 1, \ldots, N_c$ as

$$Q_n^c(\delta) = \sum_{\ell=1}^{N_c}\int_{D_c^\ell}|\zeta_n^c(t^\ell_{(c-1)^+})|^2\, dw_0.$$

Let $\zeta^c(t_{(c-1)^+}) = \zeta^c(t_c, t_{c^+})$ denote a complex-valued Gaussian process with mean zero and covariance function $R_c^\zeta(t_{(c-1)^+}, t^0_{(c-1)^+}) = [\phi_{X_c}(t_c - t_c^0) - \phi_{X_c}(t_c)\overline{\phi_{X_c}(t_c^0)}][\phi_{X_{c^+}}(t_{c^+} - t^0_{c^+}) - \phi_{X_{c^+}}(t_{c^+})\overline{\phi_{X_{c^+}}(t^0_{c^+})}]$.

By the multivariate central limit theorem, delta method and continuous mapping theorem, we have

$$\begin{pmatrix} Q_n^1(\delta) - \sum_{\ell=1}^{N_1}\int_{D_1^\ell}|\zeta^1(t^\ell)|^2\, dw_0 \\ \vdots \\ Q_n^{d-1}(\delta) - \sum_{\ell=1}^{N_{d-1}}\int_{D_{d-1}^\ell}|\zeta^{d-1}(t^\ell_{(d-2)^+})|^2\, dw_0 \end{pmatrix} \to_{\mathcal{D}} \begin{pmatrix} \sum_{\ell=1}^{N_1}\int_{D_1^\ell}|\zeta^1(t^\ell)|^2\, dw_0 \\ \vdots \\ \sum_{\ell=1}^{N_{d-1}}\int_{D_{d-1}^\ell}|\zeta^{d-1}(t^\ell_{(d-2)^+})|^2\, dw_0 \end{pmatrix}$$

as $n \to \infty$ with asymptotic mutual independence.

Thus, $Q_n^c(\delta)$, $c = 1, \ldots, d-1$ are asymptotically mutually independent.

By similar steps in Theorem 5 of Székely et al. [37], we have

$\mathrm{E}|Q_n^c(\delta) - \|\zeta_n^c(t_{(c-1)^+})\|^2_{w_0}| \to 0$, $c = 1, \ldots, d-1$ as $\epsilon, \delta \to 0$.

Hence

$$\begin{pmatrix} \|\zeta_n^1(t)\|_{w_0}^2 - Q_n^1(\delta) \\ \vdots \\ \|\zeta_n^{d-1}(t_{(d-2)^+})\|_{w_0}^2 - Q_n^{d-1}(\delta) \end{pmatrix} \to_{\mathcal{P}} \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \text{ as } \epsilon, \delta \to 0.$$

By the multivariate Slutsky's theorem, we have

$$\begin{pmatrix} \|\zeta_n^1(t)\|_{w_0}^2 \\ \vdots \\ \|\zeta_n^{d-1}(t_{(d-2)^+})\|_{w_0}^2 \end{pmatrix} \to_{\mathcal{D}} \begin{pmatrix} \|\zeta^1(t)\|_{w_0}^2 \\ \vdots \\ \|\zeta^{d-1}(t_{(d-2)^+})\|_{w_0}^2 \end{pmatrix}$$

as $\epsilon, \delta \to 0$, $n \to \infty$ with asymptotic mutual independence.

Therefore

$\|\zeta_n^c(t_{(c-1)^+})\|_{w_0}^2$, $c = 1, \ldots, d-1$ are asymptotically mutually independent.

$\sum_{c=1}^{d-1} \|\zeta_n^c(t_{(c-1)^+})\|_{w_0}^2 \to_{\mathcal{D}} \sum_{c=1}^{d-1} \|\zeta^c(t_{(c-1)^+})\|_{w_0}^2$ as $n \to \infty$.

Analogous to $\zeta_n^c(t_{(c-1)^+}), \zeta^c(t_{(c-1)^+}), R_c^{\zeta}(t_{(c-1)^+}, t_{(c-1)^+}^0)$ for $\mathcal{R}_n(\mathbf{X})$, we can define $\eta_n^c(t)$, $\eta^c(t), R_c^{\eta}(t, t^0)$ for $\mathcal{S}_n(\mathbf{X})$, and prove that $\|\eta_n^c(t)\|_{w_0}^2$, $c = 1, \ldots, d$ are asymptotically mutually independent, and $\sum_{c=1}^{d} \|\eta_n^c(t)\|_{w_0}^2 \to_{\mathcal{D}} \sum_{c=1}^{d} \|\eta_c(t)\|_{w_0}^2$ as $n \to \infty$ through the same derivations.

The only differences are that we will show $\phi_{(X_1,X_2,X_3)}^n(t_1, t_2, t_3) - \phi_{X_1}^n(t_1)\phi_{(X_2,X_3)}^n(t_2, t_3)$ and $\phi_{(X_1,X_2,X_3)}^n(s_1, s_2, s_3) - \phi_{X_2}^n(s_2)\phi_{(X_1,X_3)}^n(s_1, s_3)$ are asymptotically uncorrelated.

(ii) Under H_A:

By Theorem 4, we have

$\mathcal{R}_n \to \mathcal{R} > 0$ a.s. as $n \to \infty$.

Therefore

$n\mathcal{R}_n \to \infty$ a.s. as $n \to \infty$.

Similarly, we can prove that $n\mathcal{S}_n \to \infty$ a.s. as $n \to \infty$ through the same derivations. □

Remark. Under H_A, $\zeta_n^c(t_{(c-1)^+})$, $c = 1, \ldots, d-1$ are not asymptotically uncorrelated, and $\eta_n^c(t)$, $c = 1, \ldots, d$ are not asymptotically uncorrelated.

Complete Measure of Mutual Dependence Using Weight Function w_2

Except that $\mathcal{U}_n(\mathbf{X})$ requires the additional d-th moment condition $\mathrm{E}(|X_1|_{p_1} \dots |X_d|_{p_d}) < \infty$ to be simplified, $\mathcal{U}(X)$ is in an extremely complicated form. Even when $d = 3$, $\mathcal{U}(X)$ already has 12 different terms as follows.

$$
\begin{aligned}
\mathcal{U}(X) &= \|\phi_X(t) - \phi_{\overline{X}}(t)\|_{w_2}^2 \\
&= -\mathrm{E}|X_1 - X_1'|_{p_1}|X_2 - X_2'|_{p_2}|X_3 - X_3'|_{p_3} \\
&\quad + 2\mathrm{E}|X_1 - X_1'|_{p_1}|X_2 - X_2''|_{p_2}|X_3 - X_3'''|_{p_3} \\
&\quad - \mathrm{E}|X_1 - X_1'|_{p_1}\mathrm{E}|X_2 - X_2'|_{p_2}\mathrm{E}|X_3 - X_3'|_{p_3} \\
&\quad + \mathrm{E}|X_1 - X_1'|_{p_1}|X_2 - X_2'|_{p_2} + \mathrm{E}|X_1 - X_1'|_{p_1}|X_3 - X_3'|_{p_3} + \mathrm{E}|X_2 - X_2'|_{p_2}|X_3 - X_3'|_{p_3} \\
&\quad - 2\mathrm{E}|X_1 - X_1'|_{p_1}|X_2 - X_2''|_{p_2} - 2\mathrm{E}|X_1 - X_1'|_{p_1}|X_3 - X_3'''|_{p_3} - 2\mathrm{E}|X_2 - X_2''|_{p_2}|X_3 - X_3'''|_{p_3} \\
&\quad + \mathrm{E}|X_1 - X_1'|_{p_1}\mathrm{E}|X_2 - X_2'|_{p_2} + \mathrm{E}|X_1 - X_1'|_{p_1}\mathrm{E}|X_3 - X_3'|_{p_3} + \mathrm{E}|X_2 - X_2'|_{p_2}\mathrm{E}|X_3 - X_3'|_{p_3} \\
&= -\mathrm{E}|X_1 - X_1'|_{p_1}|X_2 - X_2'|_{p_2}|X_3 - X_3'|_{p_3} \\
&\quad + 2\mathrm{E}|X_1 - X_1'|_{p_1}|X_2 - X_2''|_{p_2}|X_3 - X_3'''|_{p_3} \\
&\quad - \mathrm{E}|X_1 - X_1'|_{p_1}\mathrm{E}|X_2 - X_2'|_{p_2}\mathrm{E}|X_3 - X_3'|_{p_3} \\
&\quad + \sum_{1 \le i < j \le 3} \mathrm{E}|X_i - X_i'|_{p_i}|X_j - X_j'|_{p_j} \\
&\quad - 2\sum_{1 \le i < j \le 3} \mathrm{E}|X_i - X_i'|_{p_i}|X_j - X_j''|_{p_j} \\
&\quad + \sum_{1 \le i < j \le 3} \mathrm{E}|X_i - X_i'|_{p_i}\mathrm{E}|X_j - X_j'|_{p_j}.
\end{aligned}
$$

In general, the number of different terms in $\mathcal{U}(X)$ grows exponentially as d increases. Basically, we will see all combinations of all components in all moments as expectations.

CHAPTER 2

INDEPENDENT COMPONENT ANALYSIS VIA ENERGY-BASED AND KERNEL-BASED MUTUAL DEPENDENCE MEASURES

Independent Component Analysis via Energy-based and Kernel-based Mutual Dependence Measures

Ze Jin[*]
Department of Statistical Science
Cornell University
Ithaca, NY 14850

David S. Matteson[†]
Department of Statistical Science
Cornell University
Ithaca, NY 14850

Abstract

We apply both distance-based (Jin and Matteson, 2017) and kernel-based (Pfister et al., 2016) mutual dependence measures to independent component analysis (ICA), and generalize dCovICA (Matteson and Tsay, 2017) to MDMICA, minimizing empirical dependence measures as an objective function in both deflation and parallel manners. Solving this minimization problem, we introduce Latin hypercube sampling (LHS) (McKay et al., 2000), and a global optimization method, Bayesian optimization (BO) (Mockus, 1994) to improve the initialization of the Newton-type local optimization method. The performance of MDMICA is evaluated in various simulation studies and an image data example. When the ICA model is correct, MDMICA achieves competitive results compared to existing approaches. When the ICA model is misspecified, the estimated independent components are less mutually dependent than the observed components using MDMICA, while they are prone to be even more mutually dependent than the observed components using other approaches.

1 INTRODUCTION

Since most natural processes have multiple components, multivariate analysis is more compelling than univariate analysis. Nevertheless, multivariate analysis is considerably more complicated than univariate analysis, because it accounts for the mutual dependence between all variables. Due to the curse of dimensionality, it becomes essential to interpret multivariate data through a simplified representation via dimension reduction.

Independent component analysis (ICA) represents multivariate data by mutually independent components (ICs). Thus, linear combinations of ICs capture the structure of multivariate data even when other linear projection methods, such as principal component analysis (PCA), are not sufficient. As a classical unsupervised learning method, ICA has been developed for applications including blind source separation, feature extraction, brain imaging, etc. Hyvärinen et al. (2004) provide a comprehensive overview of ICA approaches to estimate ICs.

Let $Y = (Y_1, \ldots, Y_d)' \in \mathbb{R}^d$ be a random vector as observations. Assume that Y has a nonsingular, continuous distribution F_Y, with $\mathrm{E}(Y_j) = 0$ and $\mathrm{Var}(Y_j) < \infty$, $j = 1, \ldots, d$. Let $X = (X_1, \ldots, X_d)' \in \mathbb{R}^d$ be a random vector as ICs. In particular, the univariate components $X_1, \ldots, X_d$ are mutually independent, and at most one component X_j is Gaussian. Without loss of generality, X is assumed to be standardized such that $\mathrm{E}(X_j) = 0$ and $\mathrm{Var}(X_j) = 1$, $j = 1, \ldots, d$. A linear latent factor model to estimate X from Y is given by

$$Y = MX,$$

where $M \in \mathbb{R}^{d \times d}$ is a nonsingular mixing matrix.

Prewhitened random variables are uncorrelated and thus more convenient to work with from both practical and theoretical perspectives. Let $\Sigma_Y = \mathrm{Cov}(Y)$ be the covariance matrix of Y, and $H = \Sigma_Y^{-1/2}$ be an uncorrelating matrix. Let $Z = HY = (Z_1, \ldots, Z_d)' \in \mathbb{R}^d$ be a random vector as uncorrelated observations, such that $\Sigma_Z = \mathrm{Cov}(Z) = I_d$, the $d \times d$ identity matrix. Then the relation between Z and X is

$$X = M^{-1}Y = M^{-1}H^{-1}Z \triangleq WZ, \qquad (1)$$

where $W = M^{-1}H^{-1} \in \mathbb{R}^{d \times d}$ is a nonsingular unmixing matrix. Given that Z are uncorrelated, W is an

[*]Corresponding author. Email address: zj58@cornell.edu.

[†]Research support from an NSF Award (DMS-1455172), a Xerox PARC Faculty Research Award, and Cornell University Atkinson Center for a Sustainable Future (AVF-2017).

orthogonal matrix, with $d(d-1)/2$ free elements rather than d^2. We aim to simultaneously estimate W and X, such that the components of X satisfy the assumption of mutual independence.

Many popular ICA approaches minimize the mutual information or maximize the non-Gaussianity of the estimated components under the constraint that they are uncorrelated. Examples include the fourth-moment matrix diagonalization of FOBI (Cardoso, 1989) and JADE (Cardoso and Souloumiac, 1993), the information criterion of Infomax (Bell and Sejnowski, 1995), the maximum negentropy of FastICA (Hyvärinen and Oja, 1997), and the maximum likelihood principle of ProDenICA (Hastie and Tibshirani, 2003) and Spline-LCA (Risk et al., 2017; Jin et al., 2017).

Some other ICA approaches minimize the mutual dependence between the estimated components using a specific dependence measure. While dependence measures have been extensively studied, two classes have attracted a great deal of attention. One is the distance-based energy statistics (Székely and Rizzo, 2013). Székely et al. (2007) proposed distance covariance (dCov) to measure pairwise dependence, and Jin and Matteson (2017) extended it to mutual dependence measures (MDMs). Another is the kernel-based maximum mean discrepancies (MMDs) (Gretton et al., 2007). Gretton et al. (2005) proposed Hilbert−Schmidt independence criterion (HSIC) to measure pairwise dependence, and Pfister et al. (2016) generalized it to d-variable Hilbert−Schmidt independence criterion (dHSIC) measuring mutual dependence. Sejdinovic et al. (2013) showed that these two classes of measures are equivalent in the sense that MMDs can be interpreted as energy statistics with a distance kernel, and energy statistics can be interpreted as MMDs with a negative-type semimetric.

Meanwhile, Chen and Bickel (2005) and Eriksson and Koivunen (2003) applied a characteristic function-based dependence measure to ICA, for which Jin and Matteson (2017) provided a closed-form expression as an MDM and studied its asymptotic properties. Bach and Jordan (2002) applied a kernel-based dependence measure to ICA, which was formulated as an HSIC in Gretton et al. (2005). Motivated by the properties of HSIC, Shen et al. (2007) proposed FastKICA based on a mutual dependence measure extension, which is the sum of all pairwise HSIC while its 0 value does not imply mutual independence. Inspired by the properties of dCov, Matteson and Tsay (2017) proposed dCovICA based on another mutual dependence measure extension, which is a sum of squared dCov and equals 0 if and only if mutual independence holds.

However, Matteson and Tsay (2017) only demonstrated the results of a single measure from the class of energy-statistics, using multiple values to initialize the local optimization without any comparison. Thus, in this paper, we generalize dCovICA to a new approach, MDMICA, by applying the mutual dependence measures proposed in Jin and Matteson (2017) and Pfister et al. (2016), and make two contributions as follows. First, we extend its ICA framework to accommodate mutual dependence measures from both classes of energy statistics and MMDs, and compare the performance of these measures in numerical studies. Second, we study the non-convex optimization problem when estimating ICs under this ICA framework, and investigate the improvement of using multiple values over a single value for initialization through Latin hypercube sampling, a random sampling method. In addition, we introduce a global optimization method, Bayesian optimization, to further improve the initialization of local optimization.

The rest of this paper is organized as follows. We generalize the ICA framework of dCovICA in Section 2. In Section 3, we give a brief overview of dCov and MDMs, propose the new ICA approach, MDMICA, based on MDMs, and derive its asymptotic properties. In Section 4, we introduce Latin hypercube sampling and Bayesian optimization to aid the initialization of subsequent local optimization method when estimating ICs. We present the simulation results in Section 5, and a real data example in Section 6[1]. Finally, Section 7 summarizes our work.

2 ICA FRAMEWORK

For $d \geq 2$, the group of $d \times d$ orthogonal matrices is denoted by $\mathcal{O}(d)$, and its subgroup with determinant 1 is denoted by $\mathcal{SO}(d)$. For $i \neq j$, we start with the identity matrix I_d, and substitute $\cos(\psi)$ for the (i,i) and (j,j) elements, $-\sin(\psi)$ for the (i,j) element, and $\sin(\psi)$ for the (j,i) element, then we obtain a Givens rotation matrix denoted by $G_{i,j}(\psi)$.

Let $\theta = \{\theta_{i,j} : 1 \leq i < j \leq d\}$ denote a vector of rotation angles with length $p = d(d-1)/2$, and let $\theta_i = \{\theta_{i,j} : i < j \leq d\}$ such that $\theta = \{\theta_i : 1 \leq i \leq d-1\}$. Then any rotation matrix $W \in \mathcal{SO}(d)$ can be parameterized via θ as $W(\theta)$, or equivalently a product of p Givens rotation matrices determined by θ as

$$W(\theta) = G^{(d-1)}(\theta_{d-1}) \dots G^{(1)}(\theta_1),$$

where $G^{(k)}(\theta_k) = G_{k,d}(\theta_{k,d}) \dots G_{k,k+1}(\theta_{k,k+1})$ represents the rotations of the kth row with respect to all the ℓth rows, $\ell > k$.

[1] See CRAN for an accompanying R package `EDMeasure`.

Although this decomposition is not unique, the kth row of $W(\theta)$ is the same as the kth row of the partial product $G^{(k)}(\theta_k)\dots G^{(1)}(\theta_1)$. As a result, let $X(\theta) = W(\theta)Z$, we observe that the subset of angles in $\{\theta_{i,j} : 1 \le i \le k, i < j \le d\} = \{\theta_i : 1 \le i \le k\}$ fully determines the kth element of X. We define a support of θ as

$$\Theta = \left\{ \theta_{i,j} : \begin{cases} 0 \le \theta_{1,j} \le 2\pi, & \\ 0 \le \theta_{i,j} < \pi, & i \ne 1. \end{cases} \right\}, \quad (2)$$

and its subset with respect to θ_i as Θ_i. Matteson and Tsay (2011) proved that there is a unique inverse mapping of $W \in \mathcal{SO}(d)$ into $\theta \in \Theta$, which is continuous if either all elements on the main-diagonal of W are positive, or all elements of W are nonzero.

Unfortunately, the non-identification issue regarding W and X still exists because the sign and order of the components are not identifiable. Given any signed permutation matrix $P_\pm$, (1) is equivalent to

$$(P_\pm X) = P_\pm X = P_\pm W Z = (P_\pm W)Z,$$

where $P_\pm X$ and $P_\pm W$ become an alternative to X and W, as the new ICs and unmixing matrix. However, the identification up to a signed permutation is adequate in terms of modeling multivariate data by linear combinations of ICs. To make a fair comparison between different estimates, a metric invariant to the three ambiguities, scale, sign, and order of the ICs will be presented in Section 5.

Let $\mathbf{Y} \in \mathbb{R}^{n\times d}$ be an i.i.d. sample of observations from F_Y, where $\mathbf{Y}_j \in \mathbb{R}^n$ is an i.i.d. sample of observations from F_{Y_j}, $j = 1,\dots,d$. Let $\widehat{\Sigma}_{\mathbf{Y}}$ be the sample covariance matrix of $\mathbf{Y}$, and $\widehat{H} = \widehat{\Sigma}_{\mathbf{Y}}^{-1/2}$ be the estimated uncorrelating matrix. Although Σ_Y is unknown in practice, the sample covariance is a consistent estimate under the finite second-moment assumption, i.e., $\widehat{\Sigma}_{\mathbf{Y}} \xrightarrow{a.s.} \Sigma_Y$ as $n \to \infty$. Let $\widehat{\mathbf{Z}} = \mathbf{Y}\widehat{H}' \in \mathbb{R}^{n\times d}$ be the estimated uncorrelated observations, such that $\widehat{\Sigma}_{\widehat{\mathbf{Z}}} = I_d$, and $\Sigma_{\widehat{\mathbf{Z}}} \xrightarrow{a.s.} I_d$ as $n \to \infty$.

To simplify notation, we assume that $\mathbf{Z}$, an uncorrelated i.i.d. sample is given, with mean zero and unit variance. Let $\mathbf{X}(\theta) = \mathbf{Z}W(\theta)' \in \mathbb{R}^{n\times d}$ be a sample of X. Then we estimate $W(\theta)$ through θ, and define an ICA estimator as

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} f(\mathbf{X}(\theta)) = \arg\min_{\theta\in\Theta} f(\mathbf{Z}W(\theta)'), \quad (3)$$

where f is an objective function measuring the mutual dependence among $\mathbf{X}(\theta)$. Given the estimate $\widehat{\theta}$, the estimated unmixing matrix is $\widehat{W} = W(\widehat{\theta})$, and the estimated ICs are $\widehat{\mathbf{X}} = \mathbf{X}(\widehat{\theta}) = \mathbf{Z}\widehat{W}' = \mathbf{Z}W(\widehat{\theta})'$.

3 APPLYING MDM TO ICA

We reduce the estimation of ICs to the problem of choosing the function f in (3), which is expected to be a measure of mutual dependence. Following Matteson and Tsay (2017), we primarily focus on distance-based energy statistics because of their compact representations as expectations of pairwise Euclidean distances, while all the results can be easily extended to kernel-based MMDs according to the equivalence between these two classes in Sejdinovic et al. (2013).

We use $(\cdot,\cdot,\dots,\cdot)$ to concatenate (vector) components into a vector. Let $t = (t_1,\dots,t_d), X = (X_1,\dots,X_d) \in \mathbb{R}^p$ where $t_j, X_j \in \mathbb{R}^{p_j}$, p_j is a marginal dimension, $j = 1,\dots,d$, and $p = \sum_{j=1}^d p_j$ is the total dimension. The subset of components to the right of X_c is denoted by $X_{c^+} = (X_{c+1},\dots,X_d)$, $c = 0,1,\dots,d-1$. The subset of components excluding X_c is denoted by $X_{-c} = (X_1,\dots,X_{c-1},X_{c^+})$, $c = 1,\dots,d-1$. The "X" under the assumption that $X_1,\dots,X_d$ are mutually independent is denoted by $\widetilde{X} = (\widetilde{X}_1,\dots,\widetilde{X}_d)$, where $\widetilde{X}_j \overset{\mathcal{D}}{=} X_j$, $j = 1,\dots,d$, $\widetilde{X}_1,\dots,\widetilde{X}_d$ are mutually independent, while $X, \widetilde{X}$ are independent. Let X', X'' be independent copies of X such that X', X'' have the same distribution as X, while they are all independent, i.e., $X, X', X'' \overset{i.i.d.}{\sim} F_X$, and $\widetilde{X}'$ be an independent copy of $\widetilde{X}$. The Euclidean norm of X is denoted by $|X|$. The weighted $\mathcal{L}_2$ norm $\|\cdot\|_w$ of any complex-valued function $\eta(t)$ is defined by $\|\eta(t)\|_w^2 = \int_{\mathbb{R}^p} |\eta(t)|^2 w(t)\,dt$ where $|\eta(t)|^2 = \eta(t)\overline{\eta(t)}$, $\overline{\eta(t)}$ is the complex conjugate of $\eta(t)$, and $w(t)$ is any positive weight function for which the integral exists.

Let $\mathbf{X} = \{X^k = (X_1^k,\dots,X_d^k) : k = 1,\dots,n\}$ be an i.i.d. sample from F_X, the joint distribution of X, and let $\mathbf{X}_j = \{X_j^k : k = 1,\dots,n\}$ be the corresponding i.i.d. sample from F_{X_j}, the marginal distribution of X_j, $j = 1,\dots,d$, such that $\mathbf{X} = \{\mathbf{X}_1,\dots,\mathbf{X}_d\}$. Denote the joint characteristic function of X as $\phi_X(t) = \mathrm{E}[e^{i\langle t,X\rangle}]$ and its empirical version as $\phi_X^n(t) = \frac{1}{n}\sum_{k=1}^n e^{i\langle t,X^k\rangle}$, and the joint characteristic function of $\widetilde{X}$ as $\phi_{\widetilde{X}}(t) = \prod_{j=1}^d \mathrm{E}[e^{i\langle t_j,X_j\rangle}]$, and its empirical version as $\phi_{\widetilde{X}}^n(t) = \prod_{j=1}^d (\frac{1}{n}\sum_{k=1}^n e^{i\langle t_j,X_j^k\rangle})$. In addition, a simplified empirical version of $\phi_{\widetilde{X}}(t)$ is defined by $\phi_{\widetilde{X}}^{n\star}(t) = \frac{1}{n}\sum_{k=1}^n e^{i\langle t,(X_1^k,\dots,X_d^{k+d-1})\rangle}$ to substitute $\phi_{\widetilde{X}}^n(t)$ as a simplification, where X_j^{n+k} is interpreted as X_j^k for $k > 0$.

3.1 DISTANCE COVARIANCE ($d = 2$)

Székely et al. (2007) proposed distance covariance to capture non-linear and non-monotone pairwise depen-

dence between two random vectors, i.e., $X = (X_1, X_2)$.

The nonnegative distance covariance $\mathcal{V}(X)$ is defined by

$$\begin{aligned}\mathcal{V}^2(X) &= \|\phi_X(t) - \phi_{\widetilde{X}}(t)\|^2_{w_1} \\ &= \int_{\mathbb{R}^p} |\phi_X(t) - \phi_{\widetilde{X}}(t)|^2 w_1(t)\, dt,\end{aligned}$$

where the weight $w_1(t) = (K_{p_1} K_{p_2} |t_1|^{p_1+1} |t_2|^{p_2+1})^{-1}$, $K_q = \frac{2\pi^{q/2}\Gamma(1/2)}{2\Gamma((q+1)/2)}$, and Γ is the gamma function.

An equivalence to pairwise independence is implied by the definition of $\mathcal{V}(X)$. If $\mathrm{E}|X| < \infty$, then $\mathcal{V}(X) \in [0, \infty)$, and $\mathcal{V}(X) = 0$ if and only if X_1, X_2 are pairwise independent. In addition, if $\mathrm{E}|X_1 X_2| < \infty$, $\mathcal{V}^2(X)$ can be interpreted as expectations

$$\begin{aligned}\mathcal{V}^2(X) &= \mathrm{E}|X_1 - X_1'||X_2 - X_2'| \\ &\quad + \mathrm{E}|X_1 - X_1'|\mathrm{E}|X_2 - X_2'| \\ &\quad - 2\mathrm{E}|X_1 - X_1'||X_2 - X_2''|.\end{aligned}$$

We estimate $\mathcal{V}(X)$ by replacing the characteristic functions with the empirical characteristic functions from the sample. The nonnegative empirical distance covariance $\mathcal{V}_n(\mathbf{X})$ is defined by $\mathcal{V}_n^2(\mathbf{X}) = \|\phi_X^n(t) - \phi_{\widetilde{X}}^n(t)\|^2_{w_1} = \int_{\mathbb{R}^p} |\phi_X^n(t) - \phi_{\widetilde{X}}^n(t)|^2 w_1(t)\, dt$, which can be interpreted as complete V-statistics

$$\begin{aligned}\mathcal{V}_n^2(\mathbf{X}) &= \frac{1}{n^2}\sum_{k,\ell=1}^n |X_1^k - X_1^\ell||X_2^k - X_2^\ell| \\ &\quad + \frac{1}{n^2}\sum_{k,\ell=1}^n |X_1^k - X_1^\ell| \frac{1}{n^2}\sum_{k,\ell=1}^n |X_2^k - X_2^\ell| \\ &\quad - \frac{2}{n^3}\sum_{k,\ell,m=1}^n |X_1^k - X_1^\ell||X_2^k - X_2^m|.\end{aligned}$$

Calculating $\mathcal{V}_n^2(\mathbf{X})$ via the symmetry of Euclidian distances has the time complexity $O(n^2)$. If $\mathrm{E}|X| < \infty$, then we have $\mathcal{V}_n(\mathbf{X}) \xrightarrow{a.s.} \mathcal{V}(X)$ as $n \to \infty$.

Jin and Matteson (2017) generalized distance covariance to three mutual dependence measures capturing any form of mutual dependence between multiple random vectors, which include the asymmetric, symmetric, and complete measures below.

3.2 ASYMMETRIC AND SYMMETRIC MEASURES ($d \geq 2$)

The asymmetric and symmetric measures of mutual dependence $\mathcal{R}(X), \mathcal{S}(X)$ are defined by

$$\begin{aligned}\mathcal{R}(X) &= \sum_{c=1}^{d-1} \mathcal{V}^2((X_c, X_{c^+})), \\ \mathcal{S}(X) &= \sum_{c=1}^{d} \mathcal{V}^2((X_c, X_{-c})).\end{aligned}$$

Analogous to $\mathcal{V}(X)$, if $\mathrm{E}|X| < \infty$, then $\mathcal{R}(X), \mathcal{S}(X) \in [0, \infty)$, and $\mathcal{R}(X), \mathcal{S}(X) = 0$ if and only if $X_1, \ldots, X_d$ are mutually independent.

Correspondingly, the empirical asymmetric and symmetric measures of mutual dependence $\mathcal{R}_n(\mathbf{X}), \mathcal{S}_n(\mathbf{X})$ are defined by $\mathcal{R}_n(\mathbf{X}) = \sum_{c=1}^{d-1} \mathcal{V}_n^2((\mathbf{X}_c, \mathbf{X}_{c^+}))$, $\mathcal{S}_n(\mathbf{X}) = \sum_{c=1}^{d} \mathcal{V}_n^2((\mathbf{X}_c, \mathbf{X}_{-c}))$, which can be implemented with the time complexity $O(n^2)$. If $\mathrm{E}|X| < \infty$, then we have $\mathcal{R}_n(\mathbf{X}) \xrightarrow{a.s.} \mathcal{R}(X)$ and $\mathcal{S}_n(\mathbf{X}) \xrightarrow{a.s.} \mathcal{S}(X)$ as $n \to \infty$.

3.3 COMPLETE MEASURE ($d \geq 2$)

The complete measure of mutual dependence $\mathcal{Q}(X)$ is defined by

$$\begin{aligned}\mathcal{Q}(X) &= \|\phi_X(t) - \phi_{\widetilde{X}}(t)\|^2_{w_2} \\ &= \int_{\mathbb{R}^p} |\phi_X(t) - \phi_{\widetilde{X}}(t)|^2 w_2(t)\, dt,\end{aligned}$$

where $w_2(t) = (K_p |t|^{p+1})^{-1}$, $K_q = \frac{2\pi^{q/2}\Gamma(1/2)}{2\Gamma((q+1)/2)}$, and Γ is the gamma function.

An equivalence to mutual independence is implied by the definition of $\mathcal{Q}(X)$ as well. If $\mathrm{E}|X| < \infty$, then $\mathcal{Q}(X) \in [0, \infty)$, and $\mathcal{Q}(X) = 0$ if and only if $X_1, \ldots, X_d$ are mutually independent. In addition, $\mathcal{Q}(X)$ can be interpreted as expectations

$$\mathcal{Q}(X) = \mathrm{E}|X - \widetilde{X}'| + \mathrm{E}|X' - \widetilde{X}| - \mathrm{E}|X - X'| - \mathrm{E}|\widetilde{X} - \widetilde{X}'|.$$

We estimate $\mathcal{Q}(X)$ by two empirical versions. One is the empirical complete measure of mutual dependence $\mathcal{Q}_n(\mathbf{X})$, defined by $\mathcal{Q}_n(\mathbf{X}) = \|\phi_X^n(t) - \phi_{\widetilde{X}}^n(t)\|^2_{w_2} = \int_{\mathbb{R}^p} |\phi_X^n(t) - \phi_{\widetilde{X}}^n(t)|^2 w_2(t)\, dt$, which can be interpreted as complete V-statistics. We skip the details of $\mathcal{Q}_n$ and will not apply it to ICA, since it is computationally prohibitive with the time complexity $O(n^{2d})$. Another one is the simplified empirical complete measure of mutual dependence $\mathcal{Q}_n^\star(\mathbf{X})$, defined by $\mathcal{Q}_n^\star(\mathbf{X}) = \|\phi_X^n(t) - \phi_{\widetilde{X}}^{n\star}(t)\|^2_{w_2} = \int_{\mathbb{R}^p} |\phi_X^n(t) - \phi_{\widetilde{X}}^{n\star}(t)|^2 w_2(t)\, dt$,

which can be interpreted as incomplete V-statistics

$$\begin{aligned}\mathcal{Q}_n^\star(\mathbf{X}) &= \frac{2}{n^2}\sum_{k,\ell=1}^{n}|X^k-(X_1^\ell,\ldots,X_d^{\ell+d-1})| \\ &+\frac{1}{n^2}\sum_{k,\ell=1}^{n}|X^k-X^\ell| \\ &-\frac{1}{n^2}\sum_{k,\ell=1}^{n}|(X_1^k,\ldots,X_d^{k+d-1})-(X_1^\ell,\ldots,X_d^{\ell+d-1})|.\end{aligned}$$

The naive implementation of $\mathcal{Q}_n^\star(\mathbf{X})$ has the time complexity $O(n^2)$. If $\mathrm{E}|X|<\infty$, then $\mathcal{Q}_n(\mathbf{X}), \mathcal{Q}_n^\star(\mathbf{X}) \xrightarrow{a.s.} \mathcal{Q}(X)$ as $n\to\infty$.

3.4 MDMICA APPROACH AND ITS ASYMPTOTIC PROPERTIES

Inspired by the nice statistical properties of MDMs, we propose an ICA approach, MDMICA based on MDMs. To be specific, we define three MDMICA estimators, i.e., MDMICA (asy), MDMICA (sym), and MDMICA (com) by applying $f(\mathbf{X})=\mathcal{R}_n(\mathbf{X}), \mathcal{S}_n(\mathbf{X}), \mathcal{Q}_n^\star(\mathbf{X})$ in (3) respectively as

$$\widehat{\theta}_n^{\text{asy}} = \arg\min_{\theta\in\Theta}\mathcal{R}_n(\mathbf{X}(\theta)) = \arg\min_{\theta\in\Theta}\mathcal{R}_n(\mathbf{Z}W(\theta)'), \tag{4}$$

and similar expressions follow for $\widehat{\theta}_n^{\text{sym}}, \widehat{\theta}_n^{\text{com}}$. Further, we define another estimator, MDMICA (hsic), by applying dHSIC in the same way.

Since the ICA model only allows scalar components, we apply a special case of MDM to ICA where the marginal dimension $p_j=1$, $j=1,\ldots,d$, and the total dimension $p=d$. Without loss of generality, we assume that $\mathrm{E}(Y)=0$ and $\mathrm{Cov}(Y)=I_d$, and therefore $Z=Y$ and $\mathbf{Z}=\mathbf{Y}$ throughout this section. Let $\overline{\Theta}$ denote a large enough compact subset of the space Θ defined by (2). The asymptotic properties of the MDMICA estimators are derived as follows.

Theorem 1. *If Y has a nonsingular, continuous distribution F_Y with $E|Y|^2<\infty$, if there exists a unique minimizer $\theta_0\in\overline{\Theta}$ of (4), and if $W(\theta_0)$ satisfies the conditions for a unique continuous inverse to exist, then $\widehat{\theta}_n^{asy}\xrightarrow{a.s.}\theta_0$ as $n\to\infty$.*

When the ICA model is misspecified, convergence to the pseudo-true value θ_0 is obtained. Under similar conditions, $\widehat{\theta}_n^{\text{sym}}, \widehat{\theta}_n^{\text{com}}$ also converges a.s. as $n\to\infty$ due to similar arguments.

We then establish the root-n consistency of the MDMICA estimators under some regularity conditions no matter whether the ICA model holds or it is misspecified.

Theorem 2. *If the assumptions of Theorem 1 hold, and if the ICA model assumptions hold, then $|\widehat{\theta}_n^{asy}-\theta_0| = O_P(n^{-1/2})$.*

Theorem 3. *If the ICA model is misspecified but the remaining assumptions stated in Theorem 2 hold, and if $E[\frac{\partial}{\partial\theta}\mathcal{R}_n(\mathbf{X}(\theta))|_{\theta=\theta_0}] = o_P(n^{-1/2})$, where θ_0 denotes the pseudo-true value, then $|\widehat{\theta}_n^{asy}-\theta_0| = O_P(n^{-1/2})$.*

Under similar conditions, $\widehat{\theta}_n^{\text{sym}}, \widehat{\theta}_n^{\text{com}}$ are also consistent as $n\to\infty$ due to similar arguments.

The proofs of Theorem 1, 2, and 3 are similar to those of Theorem 2.1, 2.2, and Corollary 2.1 in Matteson and Tsay (2017) respectively, considering the same nature of $\mathcal{R}_n(\mathbf{X}), \mathcal{S}_n(\mathbf{X}), \mathcal{Q}_n^\star(\mathbf{X})$ as energy statistics, and replacing the empirical cumulative distribution function (ECDF) with the identity function in derivations.

4 IMPROVING INITIALIZATION OF LOCAL METHODS

In the literature, there are two primary schemes to estimate ICs with regard to how the optimization is implemented. For one, the components are extracted one at a time, known as the deflation scheme. For another, the components are extracted simultaneously, known as the parallel scheme. The deflation scheme has the advantage of lower computational cost over the parallel scheme. While the parallel scheme enjoys greater statistical efficiency, since the deflation scheme accumulates estimation uncertainty at each step in its sequential procedure.

For our ICA framework, the objective function f in (3) has $d(d-1)/2$ parameters $\theta_{i,j}\in\theta$, which can be estimated in both deflation (sequential) and parallel (joint) manners. Specifically, the deflation scheme estimates all $\theta_{i,j}\in\theta$ for each i at a time, while the parallel scheme estimates all $\theta_{i,j}\in\theta$ together at once.

In view of the special structures of associated measures, both deflation and parallel schemes are appropriate for MDMICA (asy), denoted by MDMICA (asy, def) and MDMICA (asy, par), while MDMICA (sym), MDMICA (com), and MDMICA (hsic) only fit the parallel scheme. The MDMICA algorithms for both deflation and parallel schemes are described in Alg. 1 below.

Estimating θ through (3) involves minimization of a nonconvex but locally convex objective function f, which requires initialization and iterative algorithms. The default method for MDMICA is a Newton-type local optimization method, for which we explore two ways of finding a good initialization.

The first way is to perform a random sampling method, Latin hypercube sampling (LHS) (McKay et al., 2000)

Algorithm 1 MDMICA ($\mathbf{Z}$, f)

1. Initialize θ and $W(\theta)$ via θ.
2. (deflation scheme)
   for $i = 1, \cdots, d-1$ **do**
   a. Solve $\widehat{\theta}_i = \arg\min_{\theta_i \in \Theta_i} f(\mathbf{Z}W(\theta)')$ using newton-type local optimization.
   b. Update $\theta_i \leftarrow \widehat{\theta}_i$.
   end for
2'. (parallel scheme)
   Solve $\widehat{\theta} = \arg\min_{\theta \in \Theta} f(\mathbf{Z}W(\theta)')$ using newton-type local optimization.
3. Output $\widehat{\theta} = \{\widehat{\theta}_i : 1 \le i \le d-1\}$, $\widehat{W} = W(\widehat{\theta})$, and $\widehat{\mathbf{X}} = \mathbf{Z}W(\widehat{\theta})'$.

uniformly over the space Θ to obtain a number of parameter values. Then we evaluate the objective function at each value and record the value minimizing it, which is used to initialize the subsequent local optimization algorithm. Based on our experience, the number of parameter values sampled should grow with the dimension.

The second way is to take advantage of a global optimization method, Bayesian optimization (BO) (Mockus, 1994), where the objective function f is treated as a black box. It is applicable when the function is expensive to evaluate, the derivative is unavailable, or the optimization problem is non-convex. Bayesian optimization is one of the most efficient approaches in terms of the number of function evaluations consumed, as Jones (2001), Brochu et al. (2010), Snoek et al. (2012) illustrated that it outperforms other state-of-the-art global optimization algorithms on a number of challenging problems.

Bayesian optimization models the objective with respect to the parameter values as a Gaussian process. A prior is set over the objective function and then updated with actual evaluations to get a posterior using the Bayesian technique. The utility-based selection of the next evaluation point on the objective function trades off between exploration and exploitation.

5 SIMULATION STUDIES

In this section, we evaluate the performance of our MDMICA estimators by performing simulations similar to Matteson and Tsay (2017), and compare them with the FastICA estimator, the Infomax estimator, and the JADE estimator. MDMICA (asy) is omitted because it is the same as dCovICA. Moreover, we elaborate on the implementation and error metric of ICA.

Furthermore, we try various options for each estimator. For FastICA, we evaluate three functions used to approximate negentropy in both deflation and parallel schemes, logarithm of hyperbolic cosine (logcosh), kurtosis (kur), and exponential (exp). For Infomax, we evaluate three nonlinear (squashing) functions, hyperbolic tangent (tanh), logistic (log), and extended Infomax (ext). For MDMICA (hsic), we investigate the Gaussian (gau) kernel. However, FastICA (kur) and FastICA (exp) are omitted since their performance is quite similar to that of FastICA (logcosh). Similarly, Infomax (log) and Infomax (ext) are omitted.

We simulate the ICs $\mathbf{X} \in \mathbb{R}^{n \times d}$ from eighteen distributions using `rjordan` in the R package `ProDenICA` (Hastie and Tibshirani, 2010) with sample size n and dimension d. See Figure 1 for the density functions of the eighteen distributions. Then we generate a mixing matrix $M \in \mathbb{R}^{d \times d}$ with condition number between 1 and 2 using `mixmat` in the R package `ProDenICA` (Hastie and Tibshirani, 2010), and obtain the observations $\mathbf{Y} = \mathbf{X}M'$, which are centered by their sample mean and then prewhitened by their sample covariance to obtain uncorrelated observations $\mathbf{Z} = \mathbf{Y}\widehat{H}'$. Finally, we obtain the estimate $\widehat{W}$ based on $\mathbf{Z}$ via (3), and evaluate the estimation accuracy by comparing the estimate $\widehat{W}$ to the ground truth $W_0 = (\widehat{H}M)^{-1}$. Moreover, the Newton-type local optimization is implemented by `nlm` in the R package `stats` (R Core Team, 2014), and Bayesian optimization is implemented by `mbo` in the R package `mlrMBO` (Bischl et al., 2018) with the Matérn 3/2 kernel.

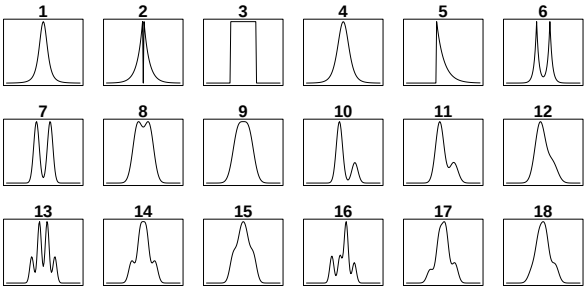

Figure 1: Density plots of the 18 distributions.

To take the uncertainty in both prewhitening the observations and estimating the ICs into account when comparing the estimates from different approaches, we use the metric MD proposed by Ilmonen et al. (2010) to measure the error between an estimate $\widehat{W}$ and the corresponding truth W_0, which is defined as

$$\mathrm{MD}(\widehat{W}, W_0) = \frac{1}{\sqrt{d-1}} \inf_{P,D} \|PD\widehat{W}W_0^{-1} - I_d\|_F,$$

where $\|\cdot\|_F$ denotes the Frobenius norm, P is a $d \times d$

permutation matrix, and D is a $d \times d$ diagonal matrix with nonzero diagonal elements. MD is invariant to the three ambiguities associated with ICA as a result of taking the infimum, for which the optimal P, D are solved by the Hungarian method (Papadimitriou and Steiglitz, 1998).

Model 1. *[Different distributions of ICs]* We sample $\mathbf{X}$ from one distribution in the eighteen distributions, with $d = 3$, $n = 1000$. We obtain $10d$ points using LHS, and select the best initial point. See Figure 2 for the error metrics of all eighteen distributions with 100 trials.

MDMICA achieves competitive results with JADE and dCovICA, and also outperforms FastICA and Infomax in most cases. MDMICA (sym) is equal and often better than dCovICA, while they have similar performance due to their similar structures. MDMICA (hsic) is equal and often better than MDMICA (com), while they have similar performance due to their similar structures. Further, MDMICA (com) and MDMICA (hsic) are less sensitive to different distributions than dCovICA and MDMICA (sym) in general. Lastly, there is no remarkable difference between the deflation and parallel schemes.

Model 2. *[Different dimensions of ICs]* We sample $\mathbf{X}$ from one distribution in the eighteen distributions, with $d \in \{2, 3, 4\}$, $n = 1000$. We pick $10d$ points using LHS, and select the best initial point. See Figure 3 for the error metrics of the 1st distribution with 100 trials.

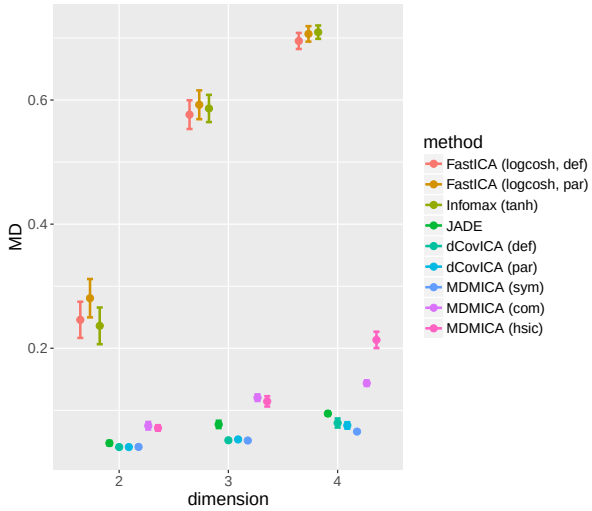

Figure 3: Error metrics (mean $\pm$ standard error) of the 1st distribution with 100 trials for Model 2.

The errors of all estimators increase as the dimension d grows. As in the previous model, JADE, dCovICA, and MDMICA have similar performance, and significantly outperform FastICA and Infomax.

Model 3. *[Different initializations of local optimization]* We sample $\mathbf{X}$ from d randomly selected distributions of the eighteen distributions, with $d = 4$, $n = 1000$. We implement three ways to select the initial point for the Newton-type local optimization method. The first way is to sample one point using LHS, and then proceed. The second way is to sample $20d$ points using LHS, and then select the point out of $20d$ with the lowest objective. The third way is to run $10d$ iterations using BO, with its initial points from $10d$ sampled points using LHS, and then select the point out of $20d$ with the lowest objective. Note that both the second way and third way run $20d$ evaluations on the objective function for a fair comparison. See Table 1 for the error metrics and objective values, and Table 2 for the computational time in initialization (LHS, BO) and local optimization (Newton-type), and total computational time of the tuple as the (4th, 11th, 12th, 18th) distributions with 100 trials.

The performance of dCovICA and MDMICA is greatly improved by selecting the best point from multiple initial points, as LHS and LHS + BO produce smaller objective values and more accurate estimates than a single point with lower mean and standard error. The reason is two-fold. First, LHS and BO offer the subsequent local optimization method better initial points in terms of lower objective, which leads to a better estimate in terms of lower objective as well. Second, a better estimate with lower objective is likely to be a better solution with lower MD, since the objective is a truly mutual dependence measure. Moreover, LHS + BO has noticeable advantage over LHS alone for MDMICA (com) and MDMICA (hsic), while it is similar to LHS alone for dCovICA (def), dCovICA (par), and MDMICA (sym).

dCovICA and MDMICA take remarkably longer computational time than the others, which makes sense because the optimization problem of dCovICA and MDMICA is especially difficult to solve, as it has $d(d-1)/2$ parameters and becomes high-dimensional quickly. This obstacle in turn motivates us to improve the local optimization by choosing a better initialization point. As LHS and BO provide better initial points for the subsequent local optimization method, the local optimization time is reduced and the total time is not necessarily longer compared to using a single initial point.

Model 4. *[Misspecified ICA model]* We sample $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})$ from one distribution in the eighteen distributions, with $n = 1000$. Let $\mathbf{Y_1} = \mathbf{X_1}$, $\mathbf{Y_2} = (\mathbf{X_2})^2$. We pick $10d$ points using LHS, and select the best initial point. See Table 3 for the results of the 1st distribution with 1 trial.

We use $\mathcal{R}_n, \mathcal{S}_n, \mathcal{Q}_n^\star$ to measure the mutual dependence between the components before (w.r.t. $\mathbf{Z}$) and after (w.r.t. $\widehat{\mathbf{X}}$) the optimization. dCovICA and MDMICA successfully decreases the mutual dependence between the com-

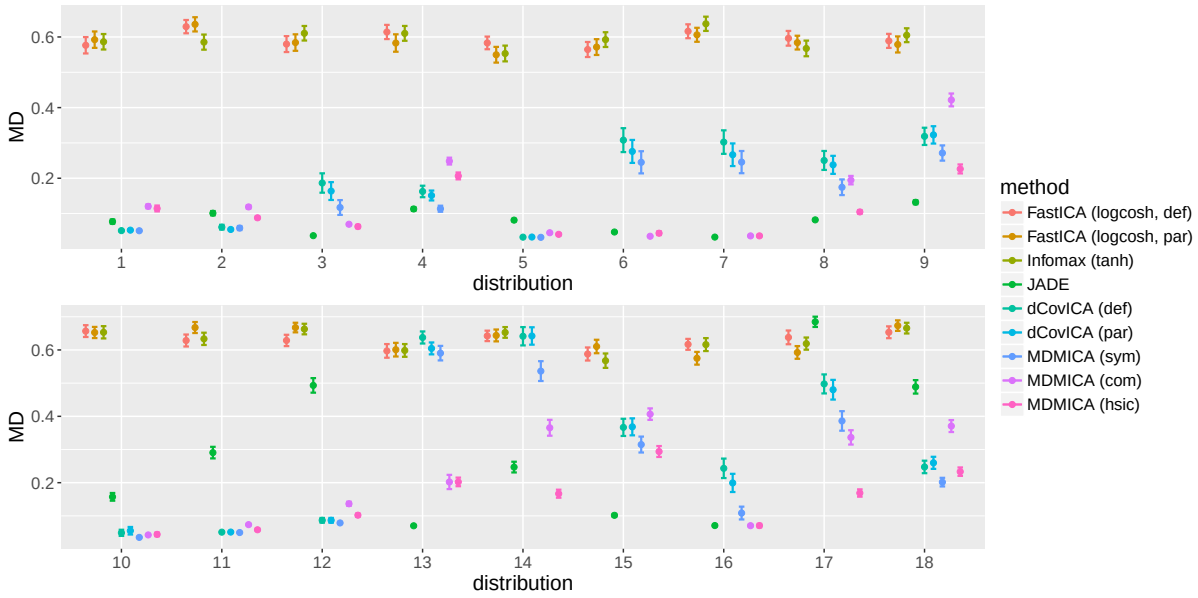

Figure 2: Error metrics (mean $\pm$ standard error) of all eighteen distributions with 100 trials for Model 1.

Table 1: Error metrics (mean $\pm$ standard error) and objective values (mean $\pm$ standard error) of the tuple as the (4th, 11th, 12th, 18th) distributions with 100 trials for Model 3.

Estimator	Initialization	MD (10^{-1})	Objective (10^{-3})
FastICA (logcosh, def)	LHS ($20d$)	6.780 ± 0.124	-
FastICA (logcosh, par)	LHS ($20d$)	6.978 ± 0.106	-
Infomax (tanh)	LHS (1)	6.861 ± 0.113	-
JADE	LHS (1)	3.992 ± 0.156	-
dCovICA (def)	LHS (1)	1.334 ± 0.105	4.090 ± 0.124
	LHS ($20d$)	1.133 ± 0.047	3.959 ± 0.047
	LHS ($10d$) + BO ($10d$)	1.128 ± 0.050	3.941 ± 0.048
dCovICA (par)	LHS (1)	1.458 ± 0.111	4.029 ± 0.060
	LHS ($20d$)	1.356 ± 0.086	3.944 ± 0.047
	LHS ($10d$) + BO ($10d$)	1.375 ± 0.099	3.963 ± 0.064
MDMICA (sym)	**LHS (1)**	$\mathbf{1.134 \pm 0.066}$	$\mathbf{6.961 \pm 0.065}$
	LHS ($20d$)	$\mathbf{1.057 \pm 0.035}$	$\mathbf{6.947 \pm 0.063}$
	LHS ($10d$) **+ BO** ($10d$)	$\mathbf{1.094 \pm 0.052}$	$\mathbf{6.952 \pm 0.065}$
MDMICA (com)	**LHS (1)**	$\mathbf{2.725 \pm 0.186}$	$\mathbf{2.037 \pm 0.093}$
	LHS ($20d$)	$\mathbf{2.064 \pm 0.097}$	$\mathbf{1.671 \pm 0.015}$
	LHS ($10d$) **+ BO** ($10d$)	$\mathbf{1.964 \pm 0.096}$	$\mathbf{1.673 \pm 0.014}$
MDMICA (hsic)	**LHS (1)**	$\mathbf{3.981 \pm 0.243}$	$\mathbf{1.385 \pm 0.091}$
	LHS ($20d$)	$\mathbf{2.521 \pm 0.152}$	$\mathbf{0.834 \pm 0.019}$
	LHS ($10d$) **+ BO** ($10d$)	$\mathbf{2.208 \pm 0.135}$	$\mathbf{0.797 \pm 0.017}$

ponents through optimization, while FastICA, Infomax, and JADE are unable to and even increase it. Therefore, ICA methods based on mutual dependence measures outperform others in reducing the mutual dependence given that the ICA model is misspecified.

6 IMAGE DATA

Fulfilling the task of unmixing vectorized images similar to Virta et al. (2016), we consider the three gray-scale images in the R package `ICS` (Nordhausen et al., 2008), depicting a cat, a forest road, and a sheep respectively. Each image is represented by a 130×130 matrix, where each element indicates the intensity value of a pixel. We standardize the three images such that the intensity val-

Table 2: Computational time (mean) in initialization (LHS, BO) and local optimization (Newton-type), and total computational time (mean) of the tuple as the (4th, 11th, 12th, 18th) distributions with 100 trials for Model 3.

Estimator	Initialization	Init Time (seconds)	Local Opt Time (seconds)	Total Time (seconds)
FastICA (logcosh, def)	LHS ($20d$)	0.13	0.03	0.16
FastICA (logcosh, par)	LHS ($20d$)	0.11	0.07	0.18
Infomax (tanh)	LHS (1)	0.00	0.05	0.05
JADE	LHS (1)	0.00	0.01	0.01
dCovICA (def)	LHS (1)	0.00	210.76	210.76
	LHS ($20d$)	221.19	174.77	395.96
	LHS ($10d$) + BO ($10d$)	207.03	174.15	381.18
dCovICA (par)	LHS (1)	0.00	910.03	910.03
	LHS ($20d$)	147.49	833.06	980.55
	LHS ($10d$) + BO ($10d$)	205.96	808.79	1014.75
MDMICA (sym)	**LHS (1)**	**0.00**	**1179.84**	**1179.84**
	LHS ($20d$)	**196.65**	**875.57**	**1072.22**
	LHS ($10d$) **+ BO** ($10d$)	**258.58**	**857.20**	**1115.78**
MDMICA (com)	**LHS (1)**	**0.00**	**92.58**	**92.58**
	LHS ($20d$)	**16.19**	**38.69**	**54.88**
	LHS ($10d$) **+ BO** ($10d$)	**78.71**	**33.39**	**112.10**
MDMICA (hsic)	**LHS (1)**	**0.00**	**267.13**	**267.13**
	LHS ($20d$)	**40.29**	**265.74**	**306.03**
	LHS ($10d$) **+ BO** ($10d$)	**106.10**	**296.03**	**402.13**

Table 3: Mutual dependence measures of observed components (before optimization, $\mathbf{Z}$) and estimated independent components (after optimization, $\widehat{\mathbf{X}}$) with 1 trial for Model 4 (misspecified ICA model).

Estimator	$\mathcal{R}_n(\mathbf{Z})$ (10^{-3})	$\mathcal{R}_n(\widehat{\mathbf{X}})$	$\mathcal{S}_n(\mathbf{Z})$ (10^{-3})	$\mathcal{S}_n(\widehat{\mathbf{X}})$	$\mathcal{Q}_n^\star(\mathbf{Z})$ (10^{-4})	$\mathcal{Q}_n^\star(\widehat{\mathbf{X}})$
FastICA (logcosh, def)		0.531		1.062		3.088
FastICA (logcosh, par)		0.588		1.176		2.786
Infomax (tanh)		0.606		1.212		3.081
JADE		1.031		2.062		3.330
dCovICA (def)	0.548	0.441	1.097	0.882	2.797	2.677
dCovICA (par)		0.441		0.882		2.677
MDMICA (sym)		**0.441**		**0.882**		**2.677**
MDMICA (com)		**0.446**		**0.892**		**2.672**
MDMICA (hsic)		**0.443**		**0.887**		**2.687**

ues across all the pixels in each image have mean zero and unit variance. Then we vectorize each image into a vector of length 130^2, and combine the vectors from all three images as a matrix $\mathbf{X}$, with $d = 3$, $n = 130^2$.

We use `mixmat` in the R package `ProDenICA` (Hastie and Tibshirani, 2010) again to generate a mixing matrix $A \in \mathbb{R}^{p\times p}$, and mix the three images to obtain the observations $\mathbf{Y} = \mathbf{X}A^T$, which are centered by their sample mean, then prewhitened by their sample covariance to obtain uncorrelated observations $\mathbf{Z} = \mathbf{Y}\widehat{H}^T$.

We estimate the intensity values $\widehat{\mathbf{S}}$ initialized from $10d$ points using LHS. See Figures 4 for the recovered images, where the Euclidean norm of vectorized errors is computed to evaluate the estimation accuracy. Indicated by the estimated images and errors, dCovICA and MDMICA outperforms JADE. Moreover, MDMICA (com) achieves the best overall performance.

7 CONCLUSION

Resorting to recently proposed mutual dependence measures including MDMs in Jin and Matteson (2017) and dHSIC in Pfister et al. (2016), we generalize dCovICA in Matteson and Tsay (2017) to a new ICA approach, MDMICA, taking empirical dependence measures as an objective function for the estimation of ICs. In addition, we study the asymptotic properties of MDMICA.

When solving the non-convex minimization problem to estimate ICs, we apply LHS and BO to select a better initial point for the Newton-type local optimization method.

MDMICA achieves competitive results with JADE and dCovICA, and outperforms FastICA and Infomax in numerical studies, under different distributions and dimensions of ICs. When the ICA model is misspecified, MDMICA decreases the mutual dependence between components via optimization, while other approaches cannot

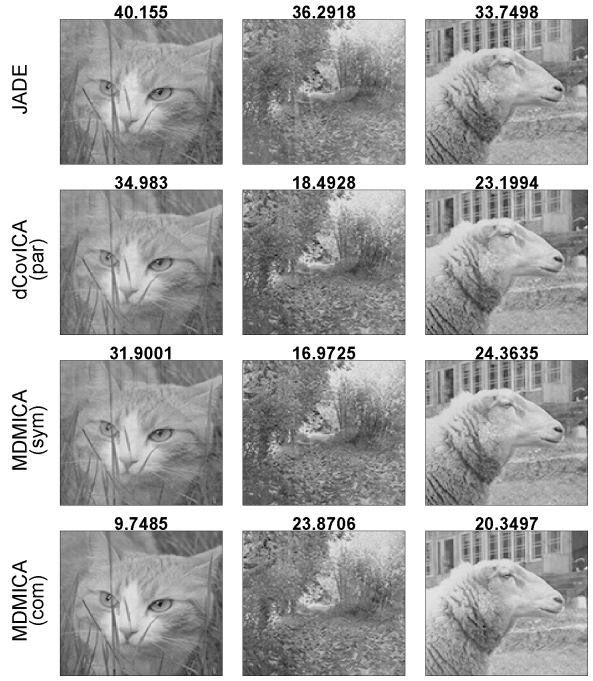

Figure 4: Recovered images with $d = 3$, $n = 130^2$ for the image data. Each value on title is the Euclidean norm of the vectorized errors of the recovered image. A signed permutation is applied to the images for illustration.

and even increase it. We illustrate the advantage of using multiple initial points from LHS and BO over a single initial point.

During the image recovery task from mixed image data, MDMICA not only nicely recovers the true images, but also achieves lower overall errors than other approaches, which demonstrates the value of MDMICA in real data applications.

References

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.

A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

B. Bischl, J. Richter, J. Bossek, D. Horn, M. Lang, and J. Thomas. *mlrMBO: Bayesian Optimization and Model-Based Optimization of Expensive Black-Box Functions*, 2018. R package version 1.1.

E. Brochu, V. M. Cora, and N. De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

J.-F. Cardoso. Source separation using higher order moments. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, pages 2109–2112. IEEE, 1989.

J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-gaussian signals. In *IEE proceedings F (radar and signal processing)*, volume 140, pages 362–370. IET, 1993.

A. Chen and P. J. Bickel. Consistent independent component analysis and prewhitening. *IEEE Transactions on Signal Processing*, 53(10):3625–3632, 2005.

J. Eriksson and V. Koivunen. Characteristic-function-based independent component analysis. *Signal Processing*, 83(10):2195–2208, 2003.

A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6 (Dec):2075–2129, 2005.

A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2007.

T. Hastie and R. Tibshirani. Independent components analysis through product density estimation. In *Advances in neural information processing systems*, pages 665–672, 2003.

T. Hastie and R. Tibshirani. *ProDenICA: Product Density Estimation for ICA using tilted Gaussian density estimates*, 2010. R package version 1.0.

A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

P. Ilmonen, K. Nordhausen, H. Oja, and E. Ollila. A new performance index for ica: properties, computation and asymptotic analysis. *Latent Variable Analysis and Signal Separation*, pages 229–236, 2010.

Z. Jin and D. S. Matteson. Generalizing distance covariance to measure and test multivariate mutual dependence. *arXiv preprint arXiv:1709.02532*, 2017.

Z. Jin, B. B. Risk, and D. S. Matteson. Optimization and testing in linear non-gaussian component analysis. *arXiv preprint arXiv:1712.08837*, 2017.

D. R. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization*, 21(4):345–383, 2001.

D. S. Matteson and R. S. Tsay. Dynamic orthogonal components for multivariate time series. *Journal of the American Statistical Association*, 106(496):1450–1463, 2011.

D. S. Matteson and R. S. Tsay. Independent component analysis via distance covariance. *Journal of the American Statistical Association*, 112:623–637, 2017.

M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1):55–61, 2000.

J. Mockus. Application of bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization*, 4(4):347–365, 1994.

K. Nordhausen, H. Oja, and D. E. Tyler. Tools for exploring multivariate data: The package ics. *Journal of Statistical Software*, 28(6):1–31, 2008.

C. H. Papadimitriou and K. Steiglitz. *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1998.

N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

B. B. Risk, D. S. Matteson, and D. Ruppert. Linear non-gaussian component analysis via maximum likelihood. *Journal of the American Statistical Association*, 2017. To appear.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.

H. Shen, S. Jegelka, and A. Gretton. Fast kernel ica using an approximate newton method. In *International Conference on Artificial Intelligence and Statistics*, pages 476–483, 2007.

J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.

J. Virta, K. Nordhausen, and H. Oja. Projection pursuit for non-gaussian independent components. *arXiv preprint arXiv:1612.05445*, 2016.

CHAPTER 3

OPTIMIZATION AND TESTING IN LINEAR NON-GAUSSIAN COMPONENT ANALYSIS

Optimization and Testing in Linear Non-Gaussian Component Analysis

Ze Jin*, Benjamin B. Risk, David S. Matteson†

June 28, 2018

Abstract

Independent component analysis (ICA) decomposes multivariate data into mutually independent components (ICs). The ICA model is subject to a constraint that at most one of these components is Gaussian, which is required for model identifiability. Linear non-Gaussian component analysis (LNGCA) generalizes the ICA model to a linear latent factor model with any number of both non-Gaussian components (signals) and Gaussian components (noise), where observations are linear combinations of independent components. Although the individual Gaussian components are not identifiable, the Gaussian subspace is identifiable. We introduce an estimator along with its optimization approach in which non-Gaussian and Gaussian components are estimated simultaneously, maximizing the discrepancy of each non-Gaussian component from Gaussianity while minimizing the discrepancy of each Gaussian component from Gaussianity. When the number of non-Gaussian components is unknown, we develop a statistical test to determine it based on resampling and the discrepancy of estimated components. Through a variety of simulation studies, we demonstrate the improvements of our estimator over competing estimators, and we illustrate the effectiveness of our test to determine the number of non-Gaussian components. Further, we apply our method to real data examples and show its practical value.

Key words: independent component analysis; multivariate analysis; hypothesis testing; subspace estimation; dimension reduction; projection pursuit

*Corresponding author. Email address: zj58@cornell.edu.

†Research support from an NSF award (DMS-1455172), a Xerox PARC Faculty Research Award, and Cornell University Atkinson Center for a Sustainable Future (AVF-2017).

1 Introduction

Independent component analysis (ICA) finds a representation of multivariate data based on mutually independent components (ICs). As an unsupervised learning method, ICA has been developed for applications including blind source separation, feature extraction, brain imaging, and many others. [1] provide an overview of ICA approaches for measuring the non-Gaussianity and estimating the ICs.

Let $Y = (Y_1, \ldots, Y_p)^T \in \mathbb{R}^p$ be a random vector of observations. Assume that Y has a nonsingular, continuous distribution F_Y, with $\mathrm{E}(Y_j) = 0$ and $\mathrm{Var}(Y_j) < \infty$, $j = 1, \ldots, p$. Let $X = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$ be a random vector of latent components. Without loss of generality, X is assumed to be standardized such that $\mathrm{E}(X_j) = 0$ and $\mathrm{Var}(X_j) = 1$, $j = 1, \ldots, p$. A static linear latent factor model to estimate the components X from the observations Y is given by

$$Y = AX,$$
$$X = A^{-1}Y \triangleq BY$$

where $A \in \mathbb{R}^{p \times p}$ is a constant, nonsingular mixing matrix, and $B \in \mathbb{R}^{p \times p}$ is a constant, nonsingular unmixing matrix.

Prewhitened random variables are uncorrelated and thus easier to work with from both practical and theoretical perspectives. Let $\Sigma_Y = \mathrm{Cov}(Y)$ be the covariance matrix of Y, and $H = \Sigma_Y^{-1/2}$ be an uncorrelating matrix. Let $Z = HY = (Z_1, \ldots, Z_p)^T \in \mathbb{R}^p$ be a random vector of uncorrelated observations, such that $\Sigma_Z = \mathrm{Cov}(Z) = I_p$, the $p \times p$ identity matrix. The ICA model further assumes that the components $X_1, \ldots, X_p$ are mutually independent, in which the number of Gaussian components is at most one. Then the relationship between

X and Z in the ICA model is

$$\begin{aligned} X &= A^{-1}Y = A^{-1}H^{-1}Z \triangleq WZ = M^T Z, \\ Z &= W^{-1}X = HAX \triangleq MX = W^T X \end{aligned} \tag{1}$$

where $W = A^{-1}H^{-1} \in \mathbb{R}^{p\times p}$ is a constant, nonsingular unmixing matrix, and $M = HA \in \mathbb{R}^{p\times p}$ is a constant, nonsingular mixing matrix. Given that Z are uncorrelated observations, W is an orthogonal matrix, and M is an orthogonal matrix as well. Thus, we have $W = M^{-1} = M^T$ and $M = W^{-1} = W^T$.

Many methods have been proposed for estimating the ICA model, including the fourth-moment diagonalization of FOBI [2] and JADE [3], the information criterion of Infomax [4], the maximum negentropy of FastICA [5], the maximum likelihood principle of ProDenICA [6], and the mutual dependence measure of dCovICA [7] and MDMICA [8]. Most of them use optimization to obtain the components such that they have maximal non-Gaussianity under the constraint that they are uncorrelated. We aim to use Z to estimate both W and X, by maximizing the non-Gaussianity of the components in X, according to a particular measure of non-Gaussianity.

To overcome the limit of the ICA model that at most one Gaussian component exists, [9] proposed the NGCA (non-Gaussian component analysis) model. Beginning with (1), the components $X \in \mathbb{R}^p$ are decomposed into signals $S \in \mathbb{R}^q$ and noise $N \in \mathbb{R}^{p-q}$, M decomposed into M_S and M_N, and W decomposed into W_S and W_N correspondingly. The components in S are assumed to be non-Gaussian, while the components in N are assumed to be Gaussian. The NGCA model further assumes that the non-Gaussian components S are independent of the Gaussian components N, the components in N are mutually independent and thus are multivariate normal, although the components in S may remain mutually dependent. Then

the relationship between X and Z in the NGCA model becomes

$$\begin{aligned} \begin{bmatrix} S \\ N \end{bmatrix} = X = WZ = \begin{bmatrix} W_S Z \\ W_N Z \end{bmatrix}, \\ Z = MX = \begin{bmatrix} M_S & M_N \end{bmatrix} \begin{bmatrix} S \\ N \end{bmatrix} = M_S S + M_N N \end{aligned} \tag{2}$$

where $M_S \in \mathbb{R}^{p\times q}$ has rank q, $M_N \in \mathbb{R}^{p\times(p-q)}$ has rank $p-q$, $W_S \in \mathbb{R}^{q\times p}$ has rank q, and $W_N \in \mathbb{R}^{(p-q)\times p}$ has rank $p-q$. The goal is to estimate the non-Gaussian subspace spanned by the rows in W_S corresponding to S, since the Gaussian subspace corresponding to N is uninteresting. [10] proved a necessary and sufficient condition for the uniqueness of the non-Gaussian subspace using projection methods. [11] developed an improved algorithm based on radial kernel functions. [12] developed theory for an approach based on characteristic functions. [13] introduced a least-squares NGCA (LSNGCA) algorithm based on least-squares estimation of log-density gradients and eigenvalue decomposition, and [14] proposed a whitening-free variant of LSNGCA. [15] developed asymptotic and bootstrap tests for the dimension of non-Gaussian subspace based on the FOBI method.

To incorporate nice characteristics from both the ICA model and NGCA model, we consider the LNGCA (linear non-Gaussian component analysis) model proposed in [16] as a special case of the NGCA model, which is the same as the the NGICA model in [17]. In the form of (2), the LNGCA model further assumes that the components $X_1, \ldots, X_p$ are mutually independent, and allows any number of both non-Gaussian components and Gaussian components among them. Similarly, we have $W = M^{-1} = M^T$ and $M = W^{-1} =$

W^T. Then the relationship between X and Z in the LNGCA model is

$$\begin{bmatrix} S \\ N \end{bmatrix} = X = WZ = \begin{bmatrix} W_S Z \\ W_N Z \end{bmatrix} = M^T Z = \begin{bmatrix} M_S^T Z \\ M_N^T Z \end{bmatrix},$$
$$Z = MX = \begin{bmatrix} M_S & M_N \end{bmatrix} \begin{bmatrix} S \\ N \end{bmatrix} = M_S S + M_N N$$

where $M_S \in \mathbb{R}^{p \times q}$ has rank q, $M_N \in \mathbb{R}^{p \times (p-q)}$ has rank $p-q$, $W_S \in \mathbb{R}^{q \times p}$ has rank q, and $W_N \in \mathbb{R}^{(p-q) \times p}$ has rank $p-q$. [16] presented a parametric LNGCA using the logistic density and a semi-parametric LNGCA using tilted Gaussians with cubic B-splines to estimate this model. [17] used projection pursuit to extract the non-Gaussian components and separate the corresponding signal and noise subspaces where the projection index is a convex combination of squared third and fourth cumulants.

In this paper, we study the LNGCA model by taking advantage of its flexibility in the number of Gaussian components, and mutual independence assumption between all components. The previous methods such as [17] and [16] only considered the discrepancy from Gaussianity for the non-Gaussian components, because with prewhitening, the Gaussian contribution to the sum of moments or model likelihood is invariant to linear transformations that preserve unit variance. Thus, an alternative framework is necessary in order to leverage the information in the Gaussian subspace. This motivates our novel objective function, which estimates the unmixing matrix W by maximizing the discrepancy from Gaussianity for the non-Gaussian components and *minimizing* the discrepancy for the Gaussian components, thereby explicitly estimating the Gaussian subspace to improve upon constrained maximum likelihood approaches. The rest of this paper is organized as follows. In Section 2, we introduce the discrepancy functions to measure the distance from Gaussianity. In Section 3, we propose a framework of LNGCA estimation given the number of non-Gaussian compo-

nents q. In Section 4, we introduce a sequence of statistical tests to determine the number of non-Gaussian components q when it is unknown. We present the simulation results in Section 5, followed by real data examples in Section 6. Finally, Section 7 is the summary of our work.

The following notations will be used throughout this paper. Let $\mathcal{O}_{a\times b}$ denote the set of $a \times b$ matrices whose columns are orthonormal. Let $\mathcal{P}^{\pm}_{a\times a}$ denote the set of $a \times a$ signed permutation matrices. Let $\|U\|_{\mathrm{F}} = \sqrt{\sum_{i,j} U_{ij}^2}$ denote the Frobenius norm of $U \in \mathbb{R}^{a\times b}$.

2 Discrepancy

2.1 Population Discrepancy Measures

In order to find the best estimate for the LNGCA model, we need a criterion to measure the discrepancy between X and its underlying assumption, i.e., S should be far from Gaussianity and N should be close to Gaussianity. Specifically, we choose a general class of functions $\mathcal{D}$ that measure the discrepancy D between each component X_j and Gaussianity.

[6] proposed the expected log-likelihood tilt function as a measure of the discrepancy from Gaussianity in the estimation of the ICA model. Suppose the density of X_j is f_j, $j = 1, \ldots, p$, and each density f_j is represented by an exponentially tilted Gaussian density

$$f_j(x_j) = \phi(x_j) e^{g_j(x_j)}$$

where ϕ is the standard univariate Gaussian density, and g_j is a smooth function. The log-tilt function g_j represents the departure from Gaussianity, and the expected log-likelihood ratio between f_j and the Gaussian density is

$$\mathrm{GPois}(X_j) = \mathrm{E}[g_j(X_j)].$$

[18, 17] proposed the use of the Jarque-Bera (JB) test statistic [19]

$$\mathrm{JB}(X_j) = \mathrm{Skew}(X_j) + \frac{1}{4}\mathrm{Kurt}(X_j)$$

to measure the discrepancy from Gaussianity in the estimation of ICA and LNGCA models, where

$$\mathrm{Skew}(X_j) = \left(\mathrm{E}[X_j^3]\right)^2,$$
$$\mathrm{Kurt}(X_j) = \left(\mathrm{E}[X_j^4] - 3\right)^2$$

are squared skewness and squared excess kurtosis. In fact, [18, 17] studied a linear combination of Skew and Kurt, i.e., $\alpha\mathrm{Skew} + (1-\alpha)\mathrm{Kurt}$, and advised the choice of $\alpha = 0.8$, which corresponds to JB. This takes deviation of both skewness and kurtosis into account, while Skew and Kurt are valid discrepancy functions as well. Notice that $\mathrm{JB}(X_j)$, $\mathrm{Skew}(X_j)$, and $\mathrm{Kurt}(X_j)$ are simplified due to standardized X_j.

2.2 Empirical Discrepancy Measures

Let $\mathbf{Y} = \{Y^i = (Y_1^i, \ldots, Y_p^i) : i = 1, \ldots, n\} \in \mathbb{R}^{n \times p}$ be an i.i.d. sample of observations from F_Y, and let $\mathbf{Y}_j = \{Y_j^i : i = 1, \ldots, n\} \in \mathbb{R}^p$ be the corresponding i.i.d. sample of observations from F_{Y_j}, $j = 1, \ldots, p$, such that $\mathbf{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_p\}$. Let $\widehat{\Sigma}_{\mathbf{Y}}$ be the sample covariance matrix of $\mathbf{Y}$, and $\widehat{H} = \widehat{\Sigma}_{\mathbf{Y}}^{-1/2}$ be the estimated uncorrelating matrix. Although the covariance Σ_Y is unknown in practice, the sample covariance $\widehat{\Sigma}_{\mathbf{Y}}$ is a consistent estimate under the assumption of finite second-moment. Let $\widehat{\mathbf{Z}} = \mathbf{Y}\widehat{H}^T \in \mathbb{R}^{n \times d}$ be the estimated uncorrelated observations, such that $\widehat{\Sigma}_{\widehat{\mathbf{Z}}} = I_d$, and $\Sigma_{\widehat{\mathbf{Z}}} \xrightarrow{a.s.} I_d$ as $n \to \infty$.

To simplify notation, we assume that $\mathbf{Z}$, an uncorrelated i.i.d. sample is given with mean zero and unit variance. Let $\mathbf{X} = \{X^i = (X_1^i, \ldots, X_p^i) : i = 1, \ldots, n\} = [\mathbf{S}, \mathbf{N}] = \mathbf{Z}W^T \in \mathbb{R}^{n \times p}$ be the sample of X, where $\mathbf{S} \in \mathbb{R}^{n \times q}$ and $\mathbf{N} \in \mathbb{R}^{n \times (p-q)}$, and let $\mathbf{X}_j = \{X_j^i : i = 1, \ldots, n\} \in$

$\mathbb{R}^n$ be the sample of X_j, i.e., the jth column in $\mathbf{X}$. Similarly, we can define $\mathbf{S}_j, \mathbf{N}_j \in \mathbb{R}^n$. Notice that $\mathbf{X}_j, \mathbf{S}_j, \mathbf{N}_j$ has sample mean 0 and sample variance 1.

We obtain the empirical discrepancy $\widehat{D}$ by replacing expectations with sample averages. The empirical GPois is given by

$$\widehat{\text{GPois}}(\mathbf{X}_j) = \frac{1}{n}\sum_{i=1}^{n} \widehat{g}_j(X_j^i)$$

where $\widehat{g}_j$ is estimated by maximum penalized likelihood, maximizing the criterion

$$\sum_{j=1}^{p}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\log\phi(X_j^i) + \widehat{g}_j(X_j^i)\right] - \lambda_j\int \widehat{g}_j''^2(x)dx\right\}$$

subject to

$$\int \phi(s)e^{\widehat{g}_j(x)}\,dx = 1$$

where $\widehat{g}_j$ is estimated by a smoothing spline, and λ_j is selected by controlling the degrees of freedom of the smoothing spline, which is 6 by default in the R package `ProDenICA` [20].

The empirical JB is given by

$$\widehat{\text{JB}}(\mathbf{X}_j) = \widehat{\text{Skew}}(\mathbf{X}_j) + \frac{1}{4}\widehat{\text{Kurt}}(\mathbf{X}_j)$$

where

$$\widehat{\text{Skew}}(\mathbf{X}_j) = \left(\frac{1}{n}\sum_{k=1}^{n}(X_j^k)^3\right)^2,$$
$$\widehat{\text{Kurt}}(\mathbf{X}_j) = \left(\frac{1}{n}\sum_{k=1}^{n}(X_j^k)^4 - 3\right)^2$$

are the empirical Skew and empirical Kurt. We will see that JB (joint use of skewness and kurtosis) performs much better than either Skew (use of skewness) or Kurt (use of kurtosis)

alone in the simulations of Section 5, which was shown in [17] as well.

3 Optimization Strategy

Using $\widehat{D}$ to measure the difference between $\mathbf{X}_j$ and Gaussianity, we seek an optimal W such that $\mathbf{X}$ is most likely to fit the underlying model with independent components.

For the ICA model, a classical ICA estimator to estimate W in FastICA [5] and Pro-DenICA [6] is defined by

$$\widehat{W}^* = \underset{W \in \mathcal{O}_{p\times p}}{\arg\max} \sum_{j=1}^{p} \widehat{D}(\mathbf{X}_j).$$

We can naturally extend the ICA estimator to an LNGCA estimator given q as

$$\widehat{W}_S^{\max} = \underset{W \in \mathcal{O}_{p\times q}}{\arg\max} \sum_{j:X_j \in S} \widehat{D}(\mathbf{X}_j) = \underset{W \in \mathcal{O}_{p\times q}}{\arg\max} \sum_{j=1}^{q} \widehat{D}(\mathbf{S}_j) \tag{3}$$

which is named the max estimator, as we maximize the discrepancy between non-Gaussian components and Gaussianity. The algorithm for the max estimator is described in Alg. 1, where the fixed point algorithm is elaborated in [6]. The objective function used in Spline-LCA from [16] is the same as the max estimator when $\widehat{D}(\cdot) = \widehat{\text{GPois}}(\cdot)$, i.e., $\widehat{D}$ and $\widehat{\text{GPois}}$ are the same empirical measure, but the optimization differs, which will be explored in Section 5.

Given the estimated unmixing matrix $\widehat{W}_S^{\max}$, the estimated non-Gaussian components are $\widehat{\mathbf{S}} = \mathbf{Z}(\widehat{W}_S^{\max})^T$.

Algorithm 1 LNGCA algorithm for the max estimator

1. Initialize $W_{p\times q}$.
2. Alternate until convergence of W, using the Frobenius norm.
   (a) Given W, estimate the discrepancy $\widehat{D}(\mathbf{S}_j)$ of component $\mathbf{S}_j$ for each j.
   (b) Given $\widehat{D}(\mathbf{S}_j)$, $j = 1, \ldots, q$, perform one step of the fixed point algorithm towards finding the optimal W.

Since any rotation of a Gaussian distribution with an identity covariance matrix leads to the same Gaussian distribution, the Gaussian components N are not identifiable. However, we can benefit from estimating the Gaussian subspace for the LNGCA model, since the column space of W_N is identifiable. Taking N into account by optimizing S and N simultaneously in the objective function, we expect to recognize the Gaussian subspace, which helps shape the non-Gaussian subspace because the non-Gaussian subspace is the complement of the Gaussian subspace. Motivated by this optimization idea, we propose a new LNGCA estimator given q as

$$\widehat{W}^{\text{max-min}} = \underset{W \in \mathcal{O}_{p\times p}}{\arg\max} \left[\sum_{j:X_j \in S} \widehat{D}(\mathbf{X}_j) - \sum_{j:X_j \in N} \widehat{D}(\mathbf{X}_j) \right] = \underset{W \in \mathcal{O}_{p\times p}}{\arg\max} \left[\sum_{j=1}^{q} \widehat{D}(\mathbf{S}_j) - \sum_{j=1}^{p-q} \widehat{D}(\mathbf{N}_j) \right] \tag{4}$$

which is named the max-min estimator for the LNGCA model, as we maximize the discrepancy between non-Gaussian components and Gaussianity, and minimize the discrepancy between Gaussian components and Gaussianity simultaneously. The algorithm for the max-min estimator is described in Alg. 2, where the fixed point algorithm is elaborated in [6]. In terms of computational complexity, the max-min estimator with p components is comparable to a typical ICA estimator with p components. Thus, the max estimator is computationally more efficient because it estimates pq parameters, whereas the max-min estimator estimates p^2 parameters as it takes the Gaussian subspace into account. We would see that the max-min estimator (joint optimization of S and N) performs much better than the max estimator (optimization of S) in the simulations of Section 5.

Given the estimated unmixing matrix $\widehat{W}^{\text{max-min}}$, the estimated non-Gaussian and Gaussian components are $\widehat{\mathbf{X}} = \mathbf{Z}(\widehat{W}^{\text{max-min}})^T$. However, it is not clear which component in $\widehat{\mathbf{X}}$ belongs to $\widehat{\mathbf{S}}$ or $\widehat{\mathbf{N}}$, since $\widehat{\mathbf{S}}$ and $\widehat{\mathbf{N}}$ are obtained together instead of $\widehat{\mathbf{S}}$ only. The solution is to sort the independent components $X_1, \ldots, X_p$ by discrepancy values in decreasing order, and obtain the ordered independent components $X_{(1)}, \ldots, X_{(p)}$. Given that there are q non-

Algorithm 2 LNGCA algorithm for the max-min estimator

1. Initialize $W_{p\times p}$.
2. Alternate until convergence of W, using the Frobenius metric.
   (a) Given W, estimate the discrepancy $\widehat{D}(\mathbf{X}_j)$ of component $\mathbf{X}_j$ for each j.
   (b) Sort components by discrepancy $\widehat{D}(\mathbf{X}_j)$ in decreasing order.
   (c) Flip the sign of discrepancy $\widehat{D}(\mathbf{X}_j)$ of the last $p-q$ components.
   (d) Given $\widehat{D}(\mathbf{X}_j)$, $j=1,\ldots,p$, perform one step of the fixed point algorithm towards finding the optimal W.
3. Sort components by discrepancy $\widehat{D}(\mathbf{X}_j)$ in decreasing order.

Gaussian components, it is natural to take $S=(X_{(1)},\ldots,X_{(q)})^T$ and $N=(X_{(q+1)},\ldots,X_{(p)})^T$ based on the discrepancy function measuring non-Gaussianity. As the q non-Gaussian components in S have the q-largest discrepancy values D among $X_1,\ldots,X_p$, the estimated non-Gaussian components in $\widehat{\mathbf{S}}$ are expected to have the q-largest empirical discrepancy values $\widehat{D}$ among $\widehat{\mathbf{X}}_1,\ldots,\widehat{\mathbf{X}}_p$.

Nevertheless, we cannot sort $\mathbf{X}$ by empirical discrepancy values to determine which component in $\mathbf{X}$ belongs to $\mathbf{S}$ or $\mathbf{N}$ at the beginning, and then stick to the order throughout the iterative algorithm and conclude which component in $\widehat{\mathbf{X}}$ belongs to $\widehat{\mathbf{S}}$ or $\widehat{\mathbf{N}}$ in the end, since the optimization does depend on the initialization, and the order of components may change after each iteration. Instead, we repeatedly sort $\mathbf{X}$ by empirical discrepancy values and adaptively determine the components in $\mathbf{S}$ and $\mathbf{N}$ at the end of each iteration in Alg 2. Finally, when the algorithm converges, we sort the estimated components $\widehat{\mathbf{X}}_1,\ldots,\widehat{\mathbf{X}}_p$ by empirical discrepancy values, and obtain the ordered estimated components $\widehat{\mathbf{X}}_{(1)},\ldots,\widehat{\mathbf{X}}_{(p)}$. Then we take $\widehat{\mathbf{S}}=[\widehat{\mathbf{X}}_{(1)},\ldots,\widehat{\mathbf{X}}_{(q)}]$, and $\widehat{\mathbf{N}}=[\widehat{\mathbf{X}}_{(q+1)},\ldots,\widehat{\mathbf{X}}_{(p)}]$. Accordingly, we decompose $\widehat{W}$ into $\widehat{W}_S$ and $\widehat{W}_N$, and $\widehat{M}=\widehat{W}^T$ into $\widehat{M}_S$ and $\widehat{M}_N$.

4 Testing and Subspace Estimation

In practice, the number of non-Gaussian components q is unknown. Following the convention of ordered components with respect to non-Gaussianity, we introduce a sequence of statistical tests to decide q. The idea behind is that, for any $j' < j$, $X_{(j')}$ is more likely to be non-Gaussian than $X_{(j)}$ in terms of discrepancy value D. If there are k non-Gaussian independent components, then $X_{(1)}, \ldots, X_{(k)}$ are non-Gaussian, and $X_{(k+1)}, \ldots, X_{(p)}$ are Gaussian.

Based on this heuristic, we propose a sequence of hypotheses for searching q as

$$H_0^{(k)} : X_{(1)}, \ldots, X_{(k-1)} \text{ are non-Gaussian and } X_{(k)}, \ldots, X_{(p)} \text{ are Gaussian,}$$
$$H_A^{(k)} : X_{(1)}, \ldots, X_{(k)} \text{ are non-Gaussian}$$

which is equivalent to testing whether there are exactly $k-1$ non-Gaussian components or at least k non-Gaussian components.

Under $H_0^{(k)}$, we first run the optimization from $\mathbf{X} = \mathbf{Z}W^T$ using the max-min estimator with $q = k-1$, in which we estimate $\widehat{W}$ and $\widehat{\mathbf{X}} = [\widehat{\mathbf{X}}_{(1)}, \ldots, \widehat{\mathbf{X}}_{(p)}]$ from the sample data $\mathbf{Z}$. One thing worth mentioning is that $\widehat{\mathbf{X}}$ depends on k as the optimization depends on k, although we suppress the notation here.

Next we repeat the following resampling procedure for B times: during the bth time, we randomly generate independent Gaussian $\mathbf{G}^{(b)} = [\mathbf{G}_1^{(b)}, \ldots, \mathbf{G}_{p-k+1}^{(b)}]$ with the same number of observations as $\mathbf{Z}$, and construct pseudo components $\mathbf{X}^{(b)} = [\widehat{\mathbf{X}}_{(1)}, \ldots, \widehat{\mathbf{X}}_{(k-1)}, \mathbf{G}^{(b)}]$. Based on the estimated unmixing matrix $\widehat{W}$, we use the estimated mixing matrix $\widehat{M} = \widehat{W}^T$ to construct pseudo observations $\mathbf{Z}^{(b)} = \mathbf{X}^{(b)}\widehat{M}^T$. Then we run the optimization from $\mathbf{X}^{(b)} = \mathbf{Z}^{(b)}W^T$ using the max-min estimator with $q = k-1$, and accordingly estimate $\widehat{W}^{(b)}$ and $\widehat{\mathbf{X}}^{(b)} = [\widehat{\mathbf{X}}_{(1)}^{(b)}, \ldots, \widehat{\mathbf{X}}_{(p)}^{(b)}]$ from the pseudo data $\mathbf{Z}^{(b)}$.

In the end, we calculate an approximate p-value by comparing $\widehat{D}(\widehat{\mathbf{X}}_{(k)})$ to $\widehat{D}(\widehat{\mathbf{X}}_{(k)}^{(b)})$, or

$\sum_{j=1}^{k} \widehat{D}(\widehat{\mathbf{X}}_{(j)})$ to $\sum_{j=1}^{k} \widehat{D}(\widehat{\mathbf{X}}_{(j)}^{(b)})$ as

$$\begin{aligned} \widehat{p}_{\text{curr}} &= \frac{\#\left\{\widehat{D}(\widehat{\mathbf{X}}_{(k)}) \leq \widehat{D}(\widehat{\mathbf{X}}_{(k)}^{(b)})\right\}}{B}, \\ \widehat{p}_{\text{cumu}} &= \frac{\#\left\{\sum_{j=1}^{k} \widehat{D}(\widehat{\mathbf{X}}_{(j)}) \leq \sum_{j=1}^{k} \widehat{D}(\widehat{\mathbf{X}}_{(j)}^{(b)})\right\}}{B} \end{aligned} \tag{5}$$

which we name the current method and the cumulative method respectively.

Our test shares the resampling technique with [15]. However, there are two major differences. On the one hand, our test does not need to bootstrap on $\mathbf{X}$, and thus saves remarkable computational cost, and we will show that it accurately estimates the number of components. On the other hand, our test is more flexible on the test statistic, as it does not need to match what is used in the objective function in the optimization. The algorithm for our sequential test is summarized in Alg. 3 below.

Algorithm 3 The algorithm for the sequential test $H_0^{(k)}$

1. Estimate $\widehat{W}$ from $\mathbf{X} = \mathbf{Z}W^T$ using the max-min estimator with $q = k - 1$.
2. Estimate $\widehat{\mathbf{X}} = \mathbf{Z}\widehat{W}^T = [\widehat{\mathbf{X}}_{(1)}, \ldots, \widehat{\mathbf{X}}_{(p)}]$.
3. Repeat the procedure for B times:
   (a) Generate independent Gaussian $\mathbf{G}^{(b)} = [\mathbf{G}_1^{(b)}, \ldots, \mathbf{G}_{p-k+1}^{(b)}]$.
   (b) Construct pseudo components $\mathbf{X}^{(b)} = [\widehat{\mathbf{X}}_{(1)}, \ldots, \widehat{\mathbf{X}}_{(k-1)}, \mathbf{G}^{(b)}]$.
   (c) Construct pseudo observations $\mathbf{Z}^{(b)} = \mathbf{X}^{(b)}\widehat{M}^T = \mathbf{X}^{(b)}\widehat{W}$.
   (d) Estimate $\widehat{W}^{(b)}$ from $\mathbf{X}^{(b)} = \mathbf{Z}^{(b)}W^T$ using the max-min estimator with $q = k - 1$.
   (e) Estimate $\widehat{\mathbf{X}}^{(b)} = \mathbf{Z}^{(b)}(\widehat{W}^b)^T = [\widehat{\mathbf{X}}_{(1)}^{(b)}, \ldots, \widehat{\mathbf{X}}_{(p)}^{(b)}]$.
3. Calculate p-value using the current or cumulative method in (5).

The proposed procedure involves a sequence of tests, but the number of tests can be dramatically reduced by using a binary search. This approach quickly narrows in on the selected q because we focus on the boundary that the p-value crosses a specific significance level. As we expect no more than $\lceil \log_2 p \rceil$ tests, it makes sense to apply the Bonferroni correction. Note that even for fairly large p, the number of tests remains reasonable, e.g., $p = 1,000$ implies fewer than ten tests. Multiple testing in this setting of sequential testing

may become more problematic as the dimension of search space grows, though the sequential searching works well in the simulations of Section 5. The issue with multiple testing is an important direction for future research.

5 Simulation Study

5.1 Sub- and Super-Gaussian Densities

In this section, we evaluate the performance of the max-min estimator by performing simulations similar to [7] for the LNGCA model, and compare it to that of the max estimator using several discrepancy functions including Skew, Kurt, JB, GPois, and Spline. Moreover, we elaborate on the implementation and performance measure of the LNGCA model.

We generate the non-Gaussian independent components $\mathbf{S} \in \mathbb{R}^{n\times q}$ from eighteen distributions using `rjordan` in the R package `ProDenICA` [20] with sample size n and dimension q. See Figure 1 for the density functions of the eighteen distributions. We also generate the Gaussian independent components $\mathbf{N} \in \mathbb{R}^{n\times(p-q)}$ with sample size n and dimension $p-q$. Then $\mathbf{X} = [\mathbf{S}, \mathbf{N}]$ are the underlying components of interest. We simulate a mixing matrix $A \in \mathbb{R}^{p\times p}$ with condition number between 1 and 2 using `mixmat` in the R package `ProDenICA` [20] and obtain the observations $\mathbf{Y} = \mathbf{X}A^T$, which are centered by their sample mean, then prewhitened by their sample covariance to obtain uncorrelated observations $\mathbf{Z} = \mathbf{Y}\widehat{H}^T$. Finally, we estimate $\widehat{W}_S$ and $\widehat{M}_S = \widehat{W}_S^T$ based on $\mathbf{Z}$ via the max estimator or the max-min estimator. Therefore, $\mathbf{Z} = \mathbf{X}A^T\widehat{H}^T = \mathbf{X}(\widehat{H}A)^T$, and we evaluate the estimation accuracy by comparing the estimated unmixing matrix $\widehat{W}$ to the ground truth $W^0 = (\widehat{H}A)^{-1} = A^{-1}\widehat{H}^{-1} = B\widehat{H}^{-1}$ with respect to S, i.e., comparing $\widehat{W}_S$ to M_S^0 where $M_S^0 = (W_S^0)^T$, $W_S^0 = B_S\widehat{H}^{-1}$.

The optimization problem associated with the max estimator in (3) and the max-min estimator in (4) is non-convex, which requires the initialization step and is sensitive to

the initial point. [21] illustrated strong sensitivity to the initialization matrix in various ICA algorithms for the eighteen distributions considered in the experiments below. To mitigate the presence of local maximum, we explore two options here, one with a single initial point, and another with multiple initial points, where each initial point is generated by orthogonalizing matrices with standard Gaussian elements, and we select the best point that has the optimal objective from multiple initial points. We suggest that the number of multiple initial points m should grow with the dimension p, e.g., $m = p$.

Each method returns an estimate for the mixing matrix. To jointly measure the uncertainty associated with both prewhitening observations and estimating non-Gaussian components, we introduce an error measure to evaluate the error between $\widehat{W}_S$ and M_S^0 as

$$\min_{Q \in \mathcal{P}_{p\times p}^{\pm}} \frac{1}{\sqrt{pq}} \|\widehat{W}_S M_S^0 - Q\|_{\mathrm{F}}^2$$

which is similar to the measures in [22], [16], and [23]. The infimum above is taken such that the measure is invariant to the sign and order of components with respect to the ambiguities associated with the LNGCA model, and the optimal Q is solved by the Hungarian method [24].

We compare the max-min estimator to the max estimator with various distributions, dimensions of components, and discrepancy functions in Experiment 1 and 2 below.

Experiment 1 (Different distributions of components)**.** We sample $\mathbf{S}$ from one of the eighteen distributions with $q = 2$, $p = 4$, and $n = 1000$. See Figure 2 for the error measures of 100 trials, with both multiple initial points ($m = 4$) and a single initial point ($m = 1$).

For both multiple initial points and a single initial point, the error measure of the max-min estimator is much lower than that of the max estimator for most distributions and discrepancy functions. Therefore, the max-min estimator improves the performance of estimation over the max estimator, no matter whether a single initial point or multiple initial

points is used in optimization.

For both the max-min estimator and max estimator, the error measure with multiple initial points is much lower than that with a single initial point for most of the distributions and discrepancy functions, which illustrates the advantage of using multiple initial points over a single initial point. Moreover, the max-min estimator using multiple initial points turns out to be a powerful combination, since the error measure of the max estimator with multiple initial points can be even further reduced when replacing the max estimator with the max-min estimator.

The error measure of JB is much lower than that of Skew and Kurt for most of the distributions, which justifies the joint use of moments. In addition, GPois is equal and often better than other discrepancy functions for all the distributions, especially with multiple initial points.

Experiment 2 (Different dimensions of components)**.** We sample $\mathbf{S}$ from q randomly selected distributions of the eighteen distributions, with $q \in \{2, 4, 8, 16\}$, $p = 2q$, $n = 500q$. See Figure 3 for the error measures of 100 trials, with both multiple initial points ($m = p$) and a single initial point ($m = 1$).

As in the previous experiment, the max-min estimator also improves the performance of estimation over the max estimator, where the error measure with multiple initial points is much lower than that with a single initial point for most cases. In addition, GPois performs the best for $q = 2, 4, 8$, and JB and GPois perform similarly for $q = 16$ with the max-min estimator and multiple initial points.

We ran further experiments similar to Experiment 1 and 2 above with various sample sizes, and observed that the performance gap between the max-min estimator and the max estimator shrinks as the sample size increases. Thus, we believe that the empirical finite-sample performance gain of the max-min estimator comes from the practical optimization advantage when simultaneously optimizing the Gaussian and non-Gaussian subspaces.

Since GPois turns out to be more robust to different distributions than Spline in the simulations, and it shares the same idea with Spline, we omit the results of Spline in the following simulation experiments and data examples.

We compare the current method with the cumulative method for selecting q with various sample sizes of components, and discrepancy functions using the max-min estimator in Experiment 3 below.

Experiment 3 (Selecting q with varying n)**.** We sample $\mathbf{S}$ from q randomly selected distributions of the eighteen distributions, with $q = 2$, $p = 4$, $n \in \{2000, 4000, 8000\}$, $B = 200$. See Table 2 and 3 for the empirical size and power of 100 trials, with significance level $\alpha = 5\%$, and both multiple initial points ($m = 4$) and a single initial point ($m = 1$).

For both multiple initial points and a single initial point, the empirical power of the current method is much higher than that of the cumulative method, and both methods have empirical size around 5% or even lower, for all the sample sizes and discrepancy functions. Thus, the current method outperforms the cumulative method in testing, no matter whether a single initial point or multiple initial points is used in optimization.

For both the current method and cumulative method, the empirical size and power with multiple initial points are similar to those with a single initial point, for all the sample sizes and discrepancy functions, which implies no remarkable effect in testing from using multiple initial points or a single initial point in estimation. This suggests that the estimate of the rank of the subspace is less sensitive to initialization than estimates of the individual components.

The empirical power of JB is much higher than that of Skew and Kurt, for all the sample sizes, which again justifies the joint use of moments. In addition, GPois outperforms the other discrepancy functions, for all the sample sizes.

5.2 Image Data

Fulfilling a task of unmixing vectorized images similar to [17], we consider the three gray-scale images from the test images of Computer Vision Group at University of Granada, depicting a cameraman, a clock, and a leopard respectively[1]. Each image is represented by a 256×256 matrix, where each element indicates the intensity value of a pixel. Three noise images of the same size are also simulated with independent standard Gaussian pixels. We standardize the six images such that the intensity values across all the pixels in each image have mean zero and unit variance. Then we vectorize each image into a vector of length 256^2, and combine the vectors from all six images as a $256^2 \times 6$ matrix $\mathbf{X}$, i.e., $p = 6$, $n = 256^2$. Thus, each row of $\mathbf{X}$ contains the intensity values of a single pixel across all images, and each column of $\mathbf{X}$ contains the intensity values of a single image.

Then we simulate a mixing matrix $A \in \mathbb{R}^{p \times p}$ using `mixmat` in the R package `ProDenICA` [20], and mix the six images to obtain the observations $\mathbf{Y} = \mathbf{X}A^T$, which are centered by their sample mean, then prewhitened by their sample covariance to get uncorrelated observations $\mathbf{Z} = \mathbf{Y}\widehat{H}^T$. We aim to infer the number of true images, and then estimate the intensity values in them.

First, we run the sequential test to infer the number of true images q with $B = 200$. See Table 1 for the p-values corresponding to each k with a single initial point ($m = 1$). Both the current method and cumulative method correctly select $q = 3$ with significance level $\alpha = 5\%$, for all the discrepancy functions.

Second, we estimate the intensity values $\widehat{\mathbf{S}}$ with $q = 3$ and multiple initial points ($m = 3$). See Figures 4 and 5 for the recovered images $\widehat{\mathbf{S}}$ and error images $\widehat{\mathbf{S}} - \mathbf{S}$, where the Euclidean norm of vectorized error images is used to evaluate the estimation accuracy. The max-min estimator outperforms the max estimator for Kurt, as the max-min estimator recovers the second image, while the second image recovered by the max estimator is masked by noise,

[1]Download data at http://decsai.ugr.es/cvg/dbimagenes/g256.php.

and also the max-min estimator has much lower error than the max-min estimator in term of the first image recovered, which illustrates the advantage of the max-min estimator over the max estimator, especially when the max estimator does not perform well. For the other discrepancy functions, both the max-min estimator and max estimator nicely recover the true images. In addition, the estimation of JB is more accurate than that of Skew and Kurt, as its recovered images are mixed with less noise, indicated by both the estimated images and error images. JB and GPois have similar performance, as JB achieves the lowest error on the first image while GPois achieves the lowest error on the second image.

6 EEG Data

There are 24 subjects in the EEG data from the Human Ecology Department at Cornell University, where each subject receives 20 trials. In each trial, 128 EEG channels (3 unused) were collected with 1024 sample points for a few seconds. We study the first trial of the first subject. The data of interest is represented by a 125×1024 matrix, i.e., $p = 125$, $n = 1024$. Here, we estimate the number of non-Gaussian signals and examine their time series. Since the max-min estimator and the current method with GPois perform the best in estimation and testing of the simulations, we only use the max-min estimator and the current method with GPois in this application.

First, we conduct the sequential test to estimate the number of non-Gaussian signals q with $B = 200$. Using the binary search for $p = 125$, we expect to have at most $\lceil \log_2 125 \rceil = 7$ tests. Therefore, we correct the significance level α to 0.714% from the original level 5%. See Figure 6 for the test statistic values ($\hat{D}(\mathbf{X}_{(k)})$) and critical values at significance level $\alpha \in \{0.714\%, 5\%, 10\%\}$ (i.e., 99.286%, 95%, and 90% quantiles of $\hat{D}(\mathbf{X}_{(k)}^{(b)})$) corresponding to $k \in \{63, 94, 110, 118, 114, 112, 113\}$ chosen from the binary search with a single initial point ($m = 1$). The current method rejects the null hypothesis that there are exactly 112

components (p-value $<$ corrected α) and fails to reject the null hypothesis that there are exactly 113 non-Gaussian components (p-value $>$ corrected α), thus selecting $q = 113$.

In the meanwhile, we iterate all $k = 1, \ldots, p$ and provide the complete testing results for reference. See Figure 7 for the test statistic values and critical values at significance level $\alpha \in \{0.714\%, 5\%, 10\%\}$ corresponding to each k with a single initial point ($m = 1$). The dashed lines pinpoint where test statistic values meet with critical values, indicating that this component is assumed to be Gaussian because we cannot reject the null hypothesis.

Second, we estimate the signals $\widehat{\mathbf{S}}$ with $q = 113$ and multiple initial points ($m = 100$). See Figure 8 for the estimated signals $\widehat{\mathbf{S}}$. The max-min estimator successfully extracts meaningful first and second components, which may be artifacts related to eye movements and eye blinks at the beginning and in the middle of the trial. The 113th and 114th components are likely to be Gaussian, as they are on the boundary of the p-value = 0.714%. The 125th (last) component is fairly close to Gaussian, compared to the Gaussian noise we randomly generate with the same sample size as a reference distribution.

We ran further experiments with 5, 10, and 20 trials, and obtained 119, 122, and 125 non-Gaussian components. One caveat to concatenating multiple trials in our EEG data is that the distribution, mean, and variance of one trial is dramatically different from another trial, which may add some noise to our results when including more trials. Additionally, as the number of trials increases, the number of eyeblink and other artifacts may increase, which contributes to a higher number of non-Gaussian components.

7 Conclusion

In this paper, we study the LNGCA model as a generalization of the ICA model, which can have any number of non-Gaussian components and Gaussian components, given that all components are mutually independent. Our contributions are the following:

(1) We propose a new max-min estimator, maximizing the discrepancy of each non-Gaussian component from Gaussianity and minimizing the discrepancy of each Gaussian component from Gaussianity simultaneously. On the contrary, the existing max estimator only maximizes the discrepancy of each non-Gaussian component from Gaussianity, which has been used in the ICA model [6] and the LNGCA model [16]. Our approach may seem unintuitive because the individual Gaussian components are not identifiable. However, the Gaussian subspace is identifiable, and joint estimation of the non-Gaussian components and Gaussian components balances the non-Gaussian subspace with the Gaussian subspace. It helps shape the non-Gaussian subspace, and thus improves the accuracy of estimating the non-Gaussian components.

(2) In practice, we need to choose the number of non-Gaussian components. We introduce a sequence of statistical tests based on generating Gaussian components and ordering estimated components by empirical discrepancy, which is computationally efficient with a binary search to reduce the actual number of tests. Two methods with different test statistics are proposed, where the current method considers the discrepancy value of the component under investigation, while the cumulative method considers the total discrepancy value of all the components from the first one up to the one under investigation. Although our test shares some characteristics with that of [15], it has less computational burden with no bootstrap needed and is more flexible in choosing the test statistics.

We evaluate the performance of our methods in simulations, demonstrating that the max-min estimator outperforms the max estimator given the number of non-Gaussian components with different discrepancy functions, dimensions, and distributions of the components, no matter whether a single initial point or multiple initial points is used in optimization. When the number of non-Gaussian components is unknown, our statistical test successfully finds the correct number with different discrepancy functions, and sample sizes, where the current method is more powerful than the cumulative method.

In the task of recovering true images from mixed image data, our test determines the correct number of true images, and we illustrate the advantage of the max-min estimator over the max estimator through some discrepancy functions. Specifically, the max-min estimator nicely recovers the images while the max estimator fails using the same discrepancy function, and the estimation error of the max-min estimator is equal and sometimes lower than of the max estimator.

In the task of exploring EEG data, our test finds a large number of non-Gaussian signals, and it extracts two components as the first two non-Gaussian components that may correspond to eye-movement and eye-blink artifacts. The distributions of estimated signals tend to become more Gaussian as their empirical discrepancy values decrease. There are a large number of non-Gaussian components in this data set. In data applications, applying a preliminary data reduction step using principal component analysis (PCA) would likely remove non-Gaussian signals. This underscores the importance of a flexible estimation and testing procedure. Further, our test assumes that the Gaussian components are i.i.d. but implicitly accounts for possible dependence in the non-Gaussian components via the resampling technique. We observed that the estimated Gaussian components from the EEG data have minor serial correlation. As an improvement, we can generate similarly serially correlated Gaussian as $\mathbf{G}^{(b)}$ in our test.

There can be several directions for the future research. One is to look for a better way to address the multiple testing issue in searching a suitable q. Another one is to better understand the improvements with the max-min estimator from a theoretical perspective. Our intuition is that the contributions of the non-Gaussian components to the asymptotic variances would equal zero. Therefore, it would be great to gain additional insight into the statistical versus computational advantages of the max-min estimator. Lastly, the max-min estimator in (4) depends on the numbers of non-Gaussian and Gaussian components. An alternative to help remove this dependency is to take the average versus the total discrepancy

of the non-Gaussian and Gaussian components, respectively. We have rerun the simulations using this method and obtained similar results, and it will be an interesting topic for future work as well.

References

[1] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

[2] J-F Cardoso. Source separation using higher order moments. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, pages 2109–2112. IEEE, 1989.

[3] Jean-François Cardoso and Antoine Souloumiac. Blind beamforming for non-gaussian signals. In *IEE proceedings F (radar and signal processing)*, volume 140, pages 362–370. IET, 1993.

[4] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[5] Aapo Hyvärinen and Erkki Oja. A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7):1483–1492, 1997.

[6] Trevor Hastie and Rob Tibshirani. Independent components analysis through product density estimation. In *Advances in neural information processing systems*, pages 665–672, 2003.

[7] David S Matteson and Ruey S Tsay. Independent component analysis via distance covariance. *Journal of the American Statistical Association*, 112(518):623–637, 2017.

[8] Ze Jin and David S Matteson. Independent component analysis via energy-based and kernel-based mutual dependence measures. *arXiv preprint arXiv:1805.06639*, 2018.

[9] Gilles Blanchard, Masashi Sugiyama, Motoaki Kawanabe, Vladimir Spokoiny, and Klaus-Robert Müller. Non-gaussian component analysis: a semi-parametric framework for linear dimension reduction. In *Advances in Neural Information Processing Systems*, pages 131–138, 2006.

[10] Fabian J Theis, Motoaki Kawanabe, and Klaus-Robert Muller. Uniqueness of non-gaussianity-based dimension reduction. *IEEE Transactions on signal processing*, 59(9): 4478–4482, 2011.

[11] Motoaki Kawanabe, Masashi Sugiyama, Gilles Blanchard, and Klaus-Robert Müller. A new algorithm of non-gaussian component analysis with radial kernel functions. *Annals of the Institute of Statistical Mathematics*, 59(1):57–75, 2007.

[12] Derek Merrill Bean. *Non-gaussian component analysis*. PhD thesis, University of California, Berkeley, 2014.

[13] Hiroaki Sasaki, Gang Niu, and Masashi Sugiyama. Non-gaussian component analysis with log-density gradient estimation. In *Artificial Intelligence and Statistics*, pages 1177–1185, 2016.

[14] Hiroaki Shiino, Hiroaki Sasaki, Gang Niu, and Masashi Sugiyama. Whitening-free least-squares non-gaussian component analysis. *arXiv preprint arXiv:1603.01029*, 2016.

[15] Klaus Nordhausen, Hannu Oja, David E Tyler, and Joni Virta. Asymptotic and bootstrap tests for the dimension of the non-gaussian subspace. *IEEE Signal Processing Letters*, 24(6):887–891, 2017.

[16] Benjamin B Risk, David S Matteson, and David Ruppert. Linear non-gaussian component analysis via maximum likelihood. *Journal of the American Statistical Association*, 2017. To appear.

[17] Joni Virta, Klaus Nordhausen, and Hannu Oja. Projection pursuit for non-gaussian independent components. *arXiv preprint arXiv:1612.05445*, 2016.

[18] Joni Virta, Klaus Nordhausen, and Hannu Oja. Joint use of third and fourth cumulants in independent component analysis. *arXiv preprint arXiv:1505.02613*, 2015.

[19] Carlos M Jarque and Anil K Bera. A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, pages 163–172, 1987.

[20] T Hastie and R Tibshirani. Prodenica: Product density estimation for ica using tilted gaussian density estimates. *R package version*, 1, 2010.

[21] Benjamin B Risk, David S Matteson, David Ruppert, Ani Eloyan, and Brian S Caffo. An evaluation of independent component analyses with an application to resting-state fmri. *Biometrics*, 70(1):224–236, 2014.

[22] Pauliina Ilmonen, Klaus Nordhausen, Hannu Oja, and Esa Ollila. A new performance index for ica: properties, computation and asymptotic analysis. *Latent Variable Analysis and Signal Separation*, pages 229–236, 2010.

[23] Jari Miettinen, Klaus Nordhausen, and Sara Taskinen. Blind source separation based on joint diagonalization in r: The packages jade and bssasymp. *Journal of Statistical Software*, 76, 2017.

[24] Christos H Papadimitriou and Kenneth Steiglitz. *Combinatorial optimization: algorithms and complexity*. Prentice-Hall, Inc., 1982.

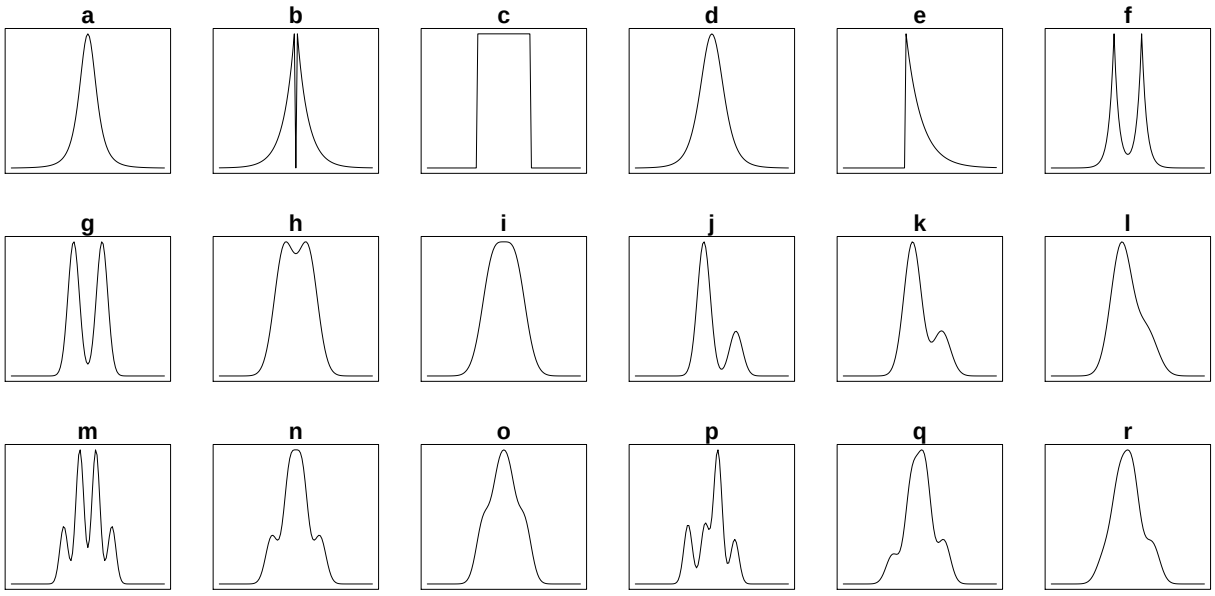

Figure 1: Density plots of the 18 distributions from `rjordan` in the R package `ProDenICA`.

Table 1: p-values of both the current method and cumulative method with $q = 3$, $p = 6$, $n = 256^2$, $B = 200$, $\alpha = 5\%$, and a single initial point ($m = 1$) in testing for the image data.

Discrepancy	Method	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
Skew	current	0.000	0.000	0.000	0.105	0.895	0.945
	cumulative	0.000	0.000	0.000	0.815	0.970	0.975
Kurt	current	0.000	0.000	0.000	0.875	0.725	0.465
	cumulative	0.000	0.000	0.055	1.000	1.000	1.000
JB	current	0.000	0.000	0.000	0.965	0.795	0.455
	cumulative	0.000	0.000	0.000	1.000	1.000	0.990
GPois	current	0.000	0.000	0.000	0.350	0.960	0.760
	cumulative	0.000	0.000	0.000	0.625	0.715	0.675

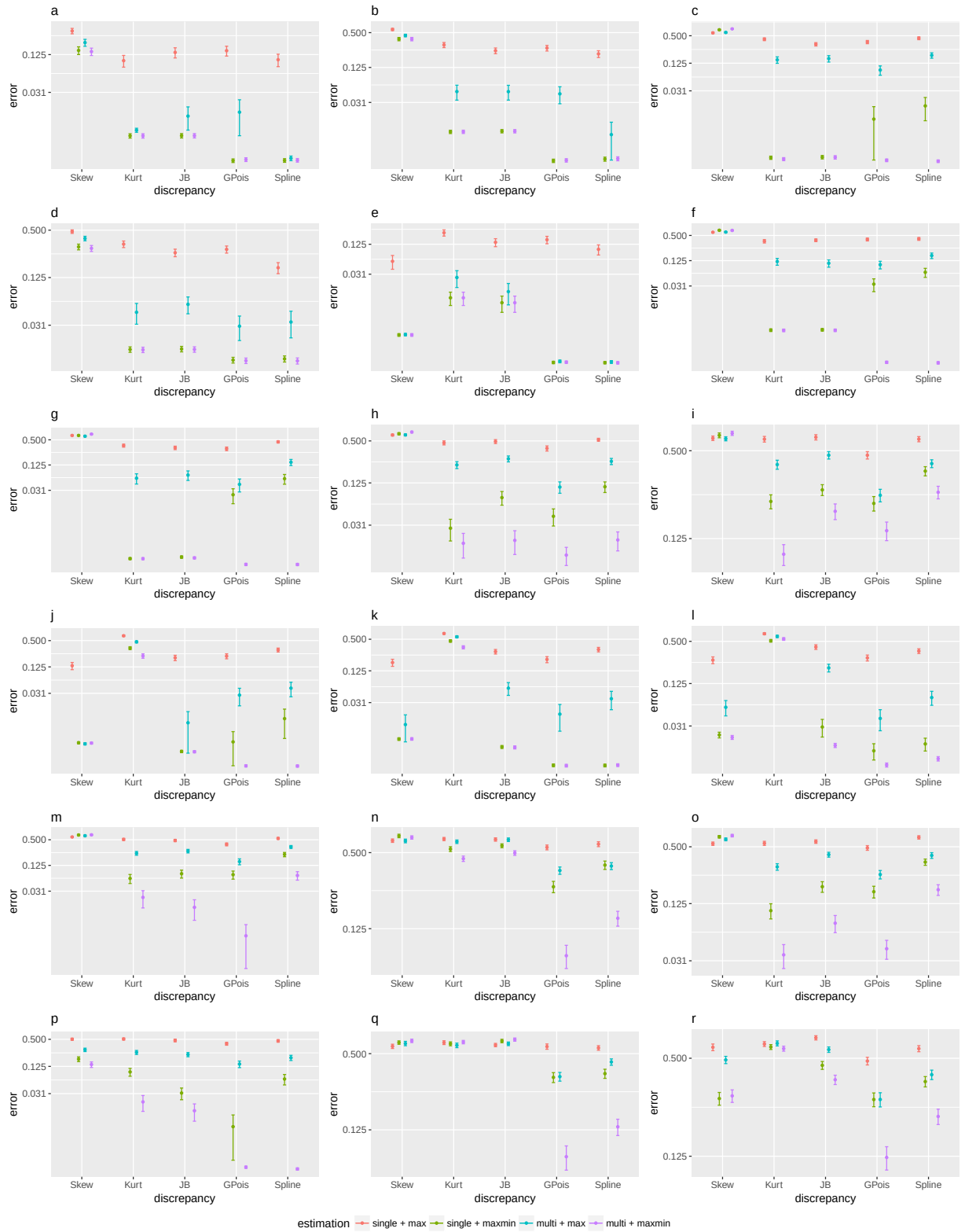

Figure 2: Error measures (mean ± standard error) in log-scale of both the max estimator and max-min estimator with $q = 2$, $p = 4$, $n = 1000$, 100 trials, and both multiple initial points ($m = 4$) and a single initial point ($m = 1$) in Experiment 1.

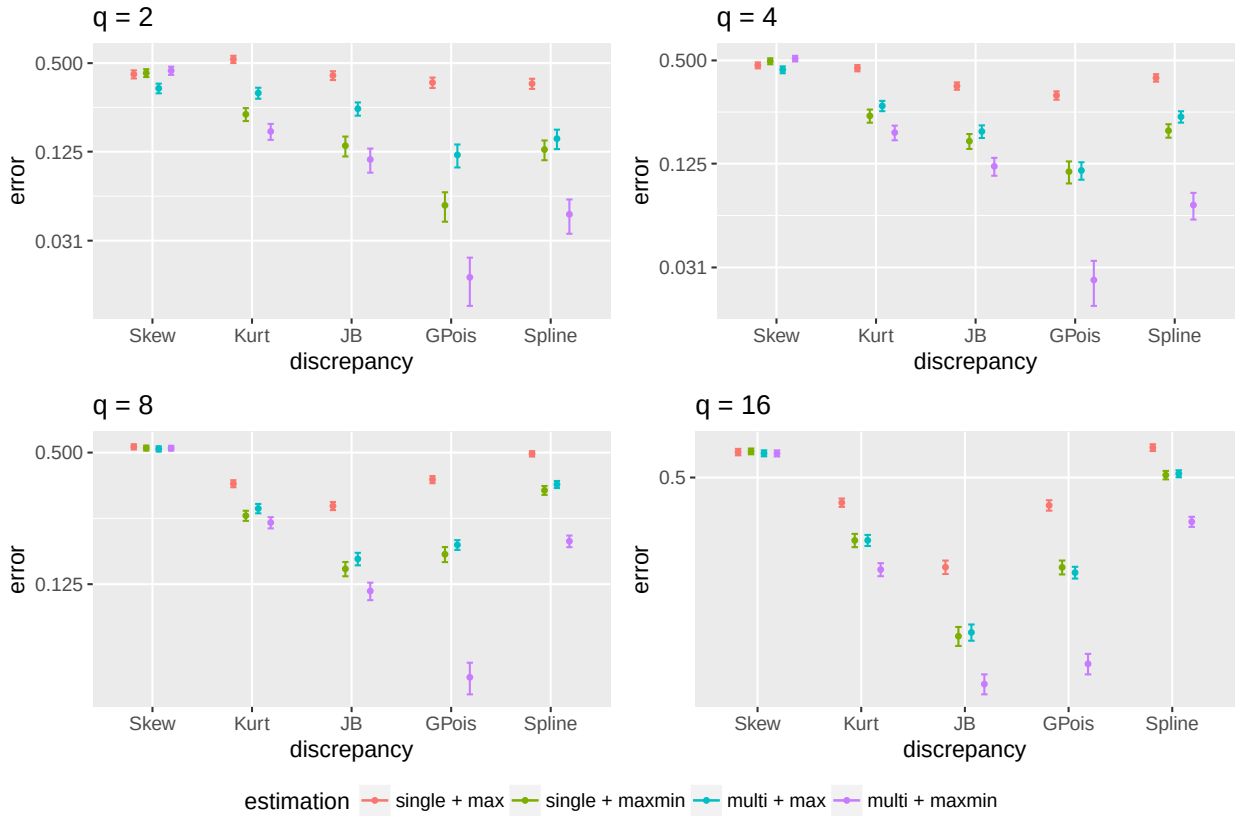

Figure 3: Error measures (mean ± standard error) in log-scale of both the max estimator and max-min estimator with $p = 2q$, $n = 500q$, 100 trials, and both multiple initial points ($m = p$) and a single initial point ($m = 1$) in Experiment 2. Error measures of a random rotation generated by orthogonalizing matrices with standard Gaussian elements are 0.537 ± 0.021, 0.674 ± 0.009, 0.801 ± 0.005, 0.916 ± 0.002 for $q = 2, 4, 8, 16$ respectively in the same setting.

Table 2: Empirical size and power of both the current method and cumulative method with $q = 2$, $p = 4$, $B = 200$, 100 trials, $\alpha = 5\%$, and a single initial point in Experiment 3.

n	Discrepancy	Method	power		size	
			$k = 1$	$k = 2$	$k = 3$	$k = 4$
2000	Skew	current	0.67	0.24	0.04	0.01
		cumulative	0.67	0.13	0.00	0.00
	Kurt	current	0.84	0.41	0.00	0.00
		cumulative	0.84	0.18	0.00	0.00
	JB	current	0.92	0.60	0.00	0.00
		cumulative	0.92	0.30	0.00	0.00
	GPois	current	1.00	0.95	0.06	0.01
		cumulative	1.00	0.94	0.00	0.00
4000	Skew	current	0.67	0.18	0.02	0.00
		cumulative	0.67	0.13	0.00	0.00
	Kurt	current	0.96	0.54	0.00	0.00
		cumulative	0.96	0.28	0.00	0.00
	JB	current	0.99	0.78	0.00	0.00
		cumulative	0.99	0.46	0.00	0.00
	GPois	current	1.00	0.99	0.05	0.03
		cumulative	1.00	0.99	0.00	0.00
8000	Skew	current	0.72	0.20	0.03	0.01
		cumulative	0.72	0.18	0.00	0.00
	Kurt	current	0.98	0.73	0.00	0.00
		cumulative	0.98	0.46	0.00	0.00
	JB	current	0.99	0.90	0.00	0.00
		cumulative	0.99	0.66	0.00	0.00
	GPois	current	1.00	1.00	0.03	0.00
		cumulative	1.00	1.00	0.00	0.00

Table 3: Empirical size and power of both the current method and cumulative method with $q = 2$, $p = 4$, $B = 200$, 100 trials, $\alpha = 5\%$, and multiple initial points in Experiment 3.

n	Discrepancy	Method	power		size	
			$k = 1$	$k = 2$	$k = 3$	$k = 4$
2000	Skew	current	0.66	0.24	0.04	0.00
		cumulative	0.66	0.13	0.00	0.00
	Kurt	current	0.86	0.41	0.00	0.00
		cumulative	0.86	0.19	0.00	0.00
	JB	current	0.94	0.61	0.00	0.00
		cumulative	0.94	0.30	0.00	0.00
	GPois	current	1.00	0.91	0.07	0.00
		cumulative	1.00	0.89	0.00	0.00
4000	Skew	current	0.66	0.19	0.01	0.00
		cumulative	0.66	0.13	0.00	0.00
	Kurt	current	0.96	0.54	0.00	0.00
		cumulative	0.96	0.28	0.00	0.00
	JB	current	0.99	0.79	0.00	0.00
		cumulative	0.99	0.47	0.00	0.00
	GPois	current	1.00	0.94	0.06	0.03
		cumulative	1.00	0.94	0.00	0.00
8000	Skew	current	0.72	0.20	0.03	0.01
		cumulative	0.72	0.18	0.00	0.00
	Kurt	current	0.97	0.73	0.00	0.00
		cumulative	0.97	0.47	0.00	0.00
	JB	current	0.99	0.90	0.00	0.00
		cumulative	0.99	0.66	0.00	0.00
	GPois	current	1.00	1.00	0.03	0.00
		cumulative	1.00	1.00	0.00	0.00

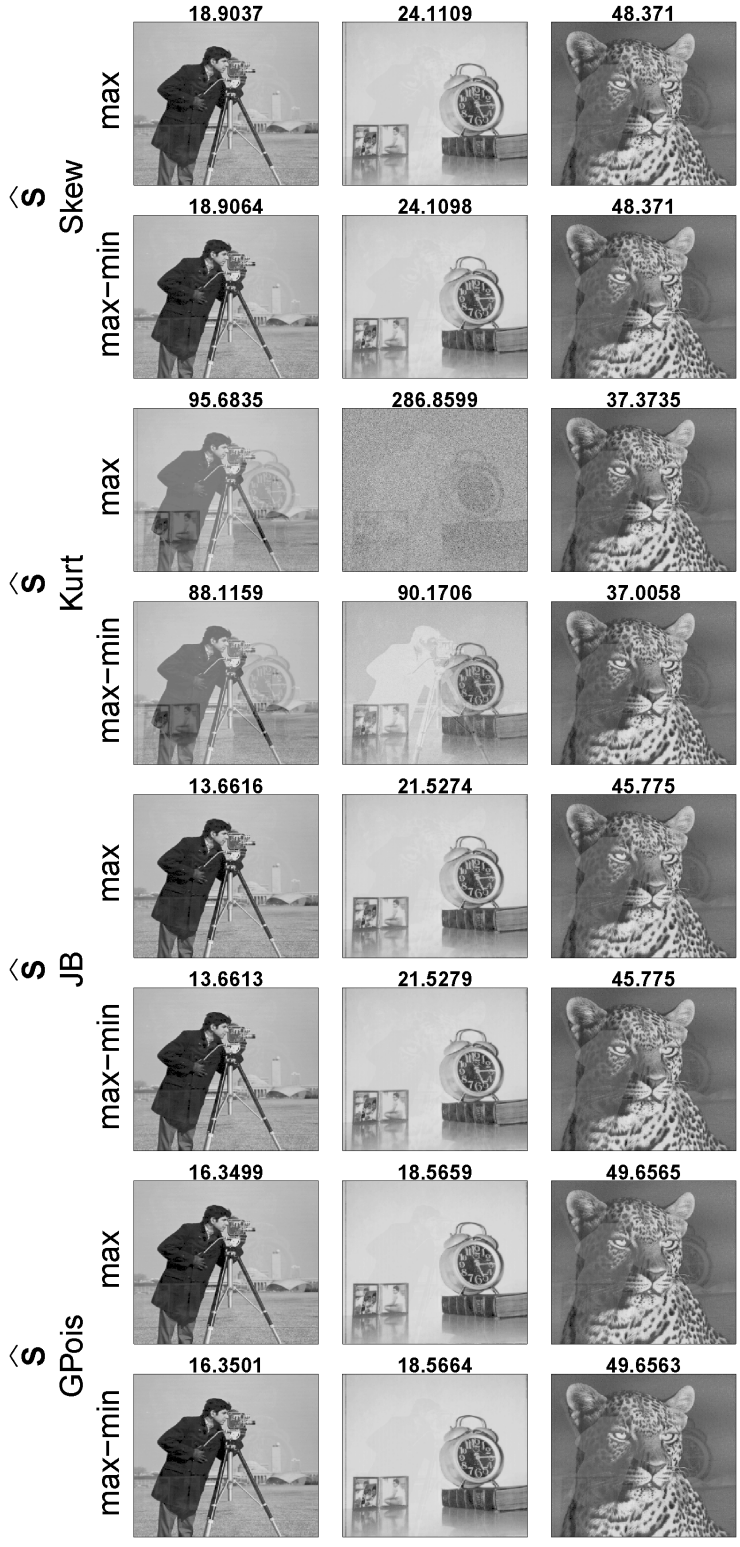

Figure 4: Recovered images of both the max estimator and max-min estimator with $q = 3$, $p = 6$, $n = 256^2$, and multiple initial points ($m = 3$) in estimation for the image data. Each value on title is the Euclidean norm of the vectorized error image corresponding to the recovered image. We apply a signed permutation to the images and modify the gray scales for illustration purpose.

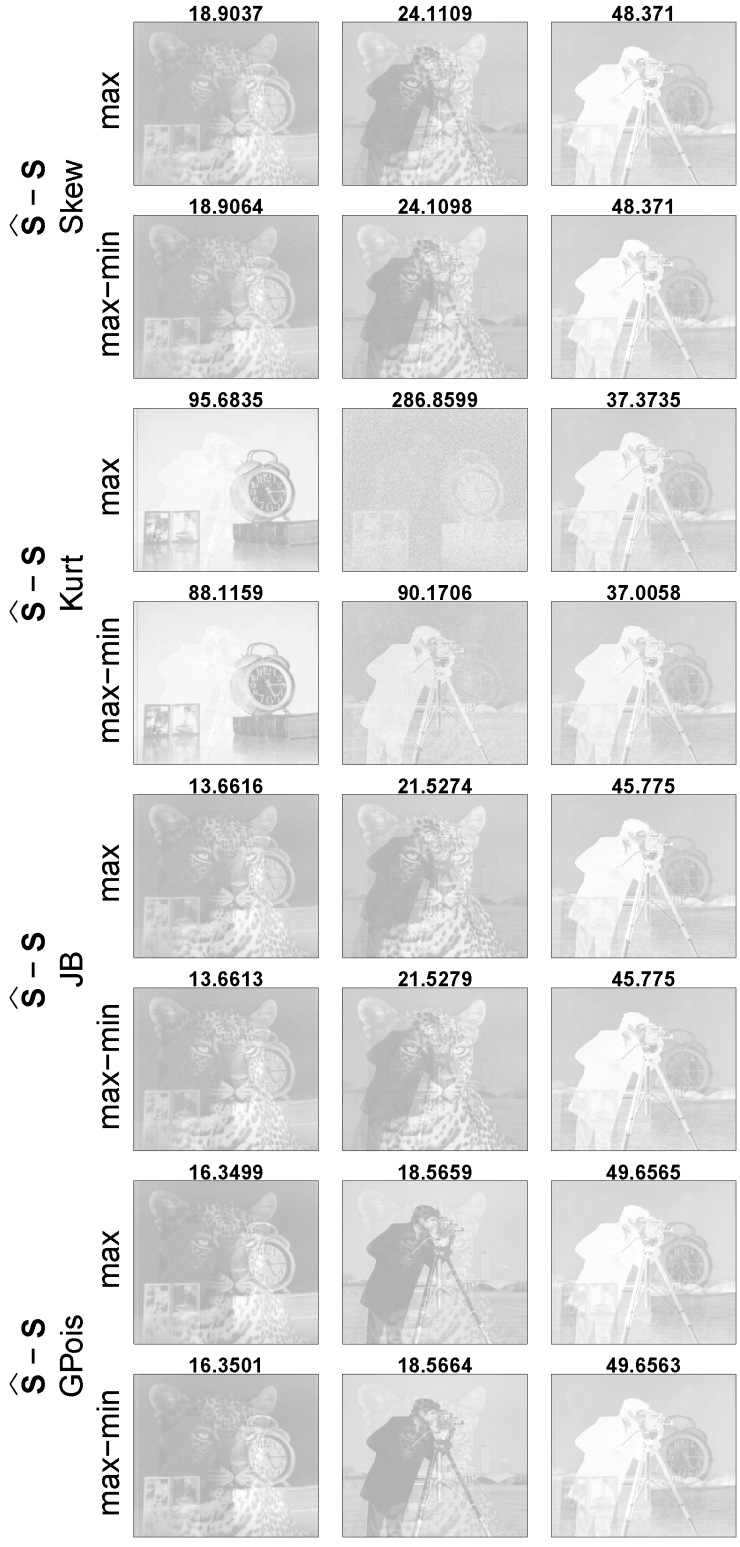

Figure 5: Error images of both the max estimator and max-min estimator with $q = 3$, $p = 6$, $n = 256^2$, and multiple initial points ($m = 3$) in estimation for the image data. Each value on title is the Euclidean norm of the vectorized error image. We apply a signed permutation to the images and modify the gray scales for illustration purpose.

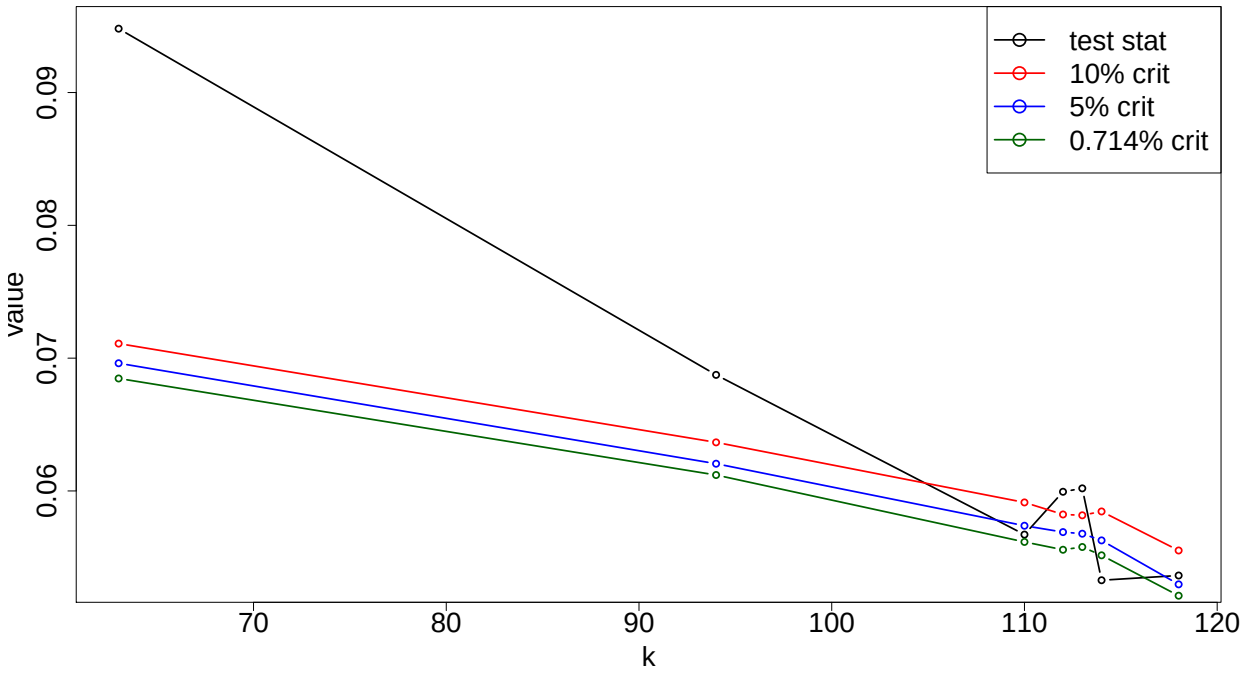

Figure 6: Test statistics and critical values of the current method for testing k from binary search with $p = 125$, $n = 1024$, $B = 200$, and a single initial point ($m = 1$) in testing for the EEG data.

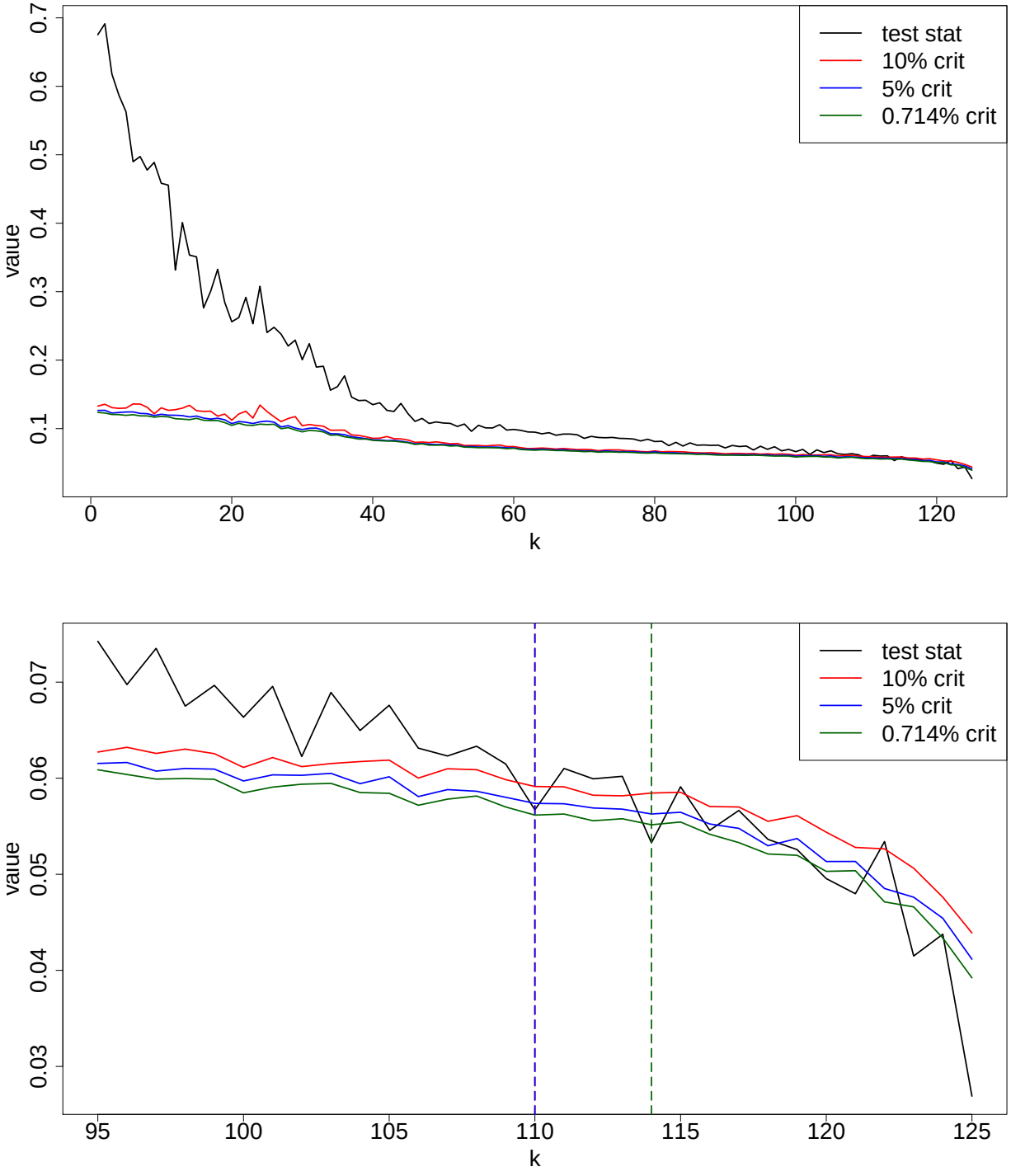

Figure 7: Test statistics and critical values of the current method for testing all k with $p = 125$, $n = 1024$, $B = 200$, and a single initial point ($m = 1$) in testing for the EEG data.

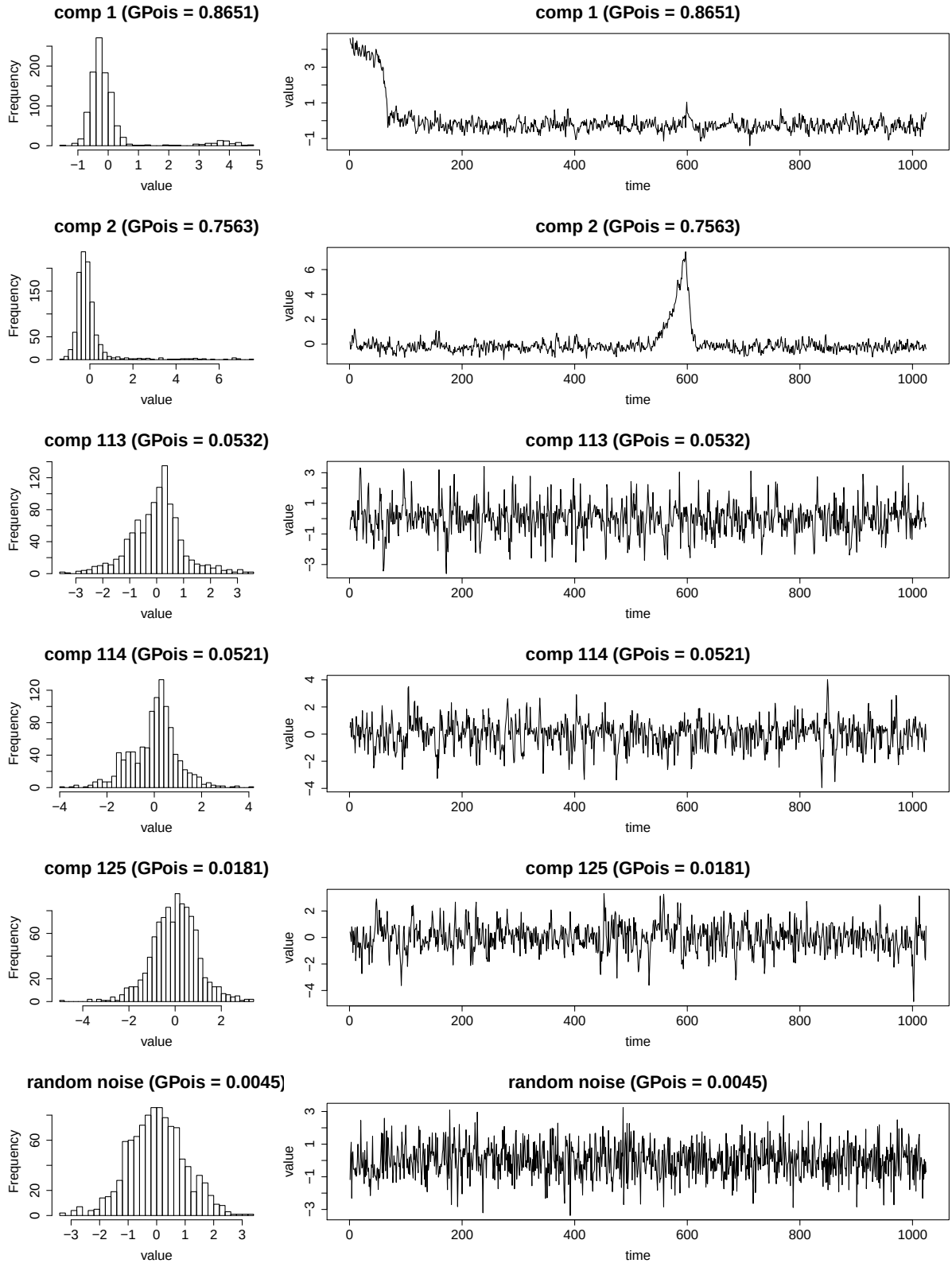

Figure 8: Estimated signals of the max-min estimator with $q = 113$, $p = 125$, $n = 1024$, and multiple initial points ($m = 100$) in estimation for the EEG data.

CHAPTER 4

TESTING FOR CONDITIONAL MEAN INDEPENDENCE WITH COVARIATES THROUGH MARTINGALE DIFFERENCE DIVERGENCE

Testing for Conditional Mean Independence with Covariates through Martingale Difference Divergence

Ze Jin[*]
Department of Statistical Science
Cornell University
Ithaca, NY 14850

Xiaohan Yan
Department of Statistical Science
Cornell University
Ithaca, NY 14850

David S. Matteson[†]
Department of Statistical Science
Cornell University
Ithaca, NY 14850

Abstract

A crucial problem in statistics is to decide whether additional variables are needed in a regression model. We propose a new multivariate test to investigate the conditional mean independence of Y given X conditioning on some known effect Z, i.e., $\mathrm{E}(Y|X,Z) = \mathrm{E}(Y|Z)$. Assuming that $\mathrm{E}(Y|Z)$ and Z are linearly related, we reformulate an equivalent notion of conditional mean independence through transformation, which is approximated in practice. We apply the martingale difference divergence (Shao and Zhang, 2014) to measure conditional mean dependence, and show that the estimation error from approximation is negligible, as it has no impact on the asymptotic distribution of the test statistic under some regularity assumptions. The implementation of our test is demonstrated by both simulations and a financial data example.

1 INTRODUCTION

Testing (conditional) dependence and conditional mean dependence plays an important role in statistics with various applications, including variable selection (Székely and Rizzo, 2014; Park et al., 2015; Zhang et al., 2015; Yan and Bien, 2018), feature screening (Li et al., 2012; Shao and Zhang, 2014; Yan et al., 2017), and graphical models (Gan et al., 2018; Li and McCormick, 2017; Li et al., 2018). Both areas attracted tremendous attention in the last two decades, as datasets have increased in size and dimension. Let $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$, $Z \in \mathbb{R}^r$ be the three random vectors of interest, and denote pairwise independence by $\perp\!\!\!\perp$.

Measures of (conditional) dependence have been extensively studied. Székely et al. (2007) proposed distance covariance (dCov) to capture the non-linear and non-monotone pairwise dependence between X and Y, and dCov = 0 if and only if pairwise independence ($X \perp\!\!\!\perp Y$) holds. Jin and Matteson (2017) extended distance covariance to mutual dependence measures (MDMs), which have been applied to independent component analysis in Jin and Matteson (2018). To capture the conditional dependence between X and Y given Z, Székely and Rizzo (2014) generalized distance covariance to partial distance covariance (pdCov), however, pdCov = 0 is not equivalent to conditional independence ($X \perp\!\!\!\perp Y|Z$), and neither one implies the other. Wang et al. (2015) extended distance covariance to conditional distance covariance (CDCov) using kernel estimators, and CDCov = 0 if and only if conditional independence holds. Under a linear assumption between X, Y and Z, Fan et al. (2015) converted testing conditional independence to testing independence, and applied distance covariance to measure the dependence of estimated variables. Moreover, inter-temporal conditional dependence is known as Granger causality in time series analysis. Hiemstra and Jones (1994), Su and White (2007), and Chen and Hong (2012) each introduced non-parametric tests for non-linear Granger causality based on conditional probabilities and characteristic functions.

Likewise, various measures of conditional mean dependence have been broadly developed as well. Testing the conditional mean independence of Y given X, i.e.,

$$H_0 : \mathrm{E}(Y|X) = \mathrm{E}(Y) \ \ a.s., \quad H_A : o.w. \qquad (1)$$

provides insight on whether X contributes to the conditional mean of Y. Shao and Zhang (2014) generalized distance covariance to martingale difference divergence (MDD), and MDD = 0 if and only if (1) holds. Testing

[*]Corresponding author. Email address: zj58@cornell.edu.

[†]Research support from an NSF Award (DMS-1455172), a Xerox PARC Faculty Research Award, and Cornell University Atkinson Center for a Sustainable Future (AVF-2017).

the conditional mean independence of Y given X conditioning on some known effect Z, i.e.,

$$H_0 : \mathrm{E}(Y|X,Z) = \mathrm{E}(Y|Z) \ \ a.s., \quad H_A : o.w. \quad (2)$$

sheds light on whether X contributes to the conditional mean of Y when taking known dependence on Z into account. Park et al. (2015) generalized martingale difference divergence to partial martingale difference divergence (pMDD), however, pMDD = 0 is not equivalent to (2). Fan and Li (1996), Lavergne and Vuong (2000), and Aït-Sahalia et al. (2001) each introduced non-parametric tests for (2) using kernel estimators of conditional expectations. Assuming a linear model between Y and (X, Z), Lan et al. (2014) generalized the classical partial F-test (Chatterjee and Hadi, 2015) to a partial covariance-based (pcov) test for (2) in the high-dimensional setting, and Tang et al. (2017) further proposed a hybrid test for (2) through finding the most predictive covariate based on both maximum-type and sum-type statistics. Conditional mean independence conditioning on lagged covariates is known as Granger causality *in mean* in time series analysis. Raïssi et al. (2011) proposed a parametric test for linear Granger causality in mean based on vector autoregressive (VAR) models, and Hong et al. (2009) introduced a non-parametric test for non-linear Granger causality in mean based on cross-correlations.

In this paper, we focus on testing conditional mean independence with covariates and develop a method to test (2) for two main reasons. As Cook and Li (2002) state, regression analysis is mostly concerned with the conditional mean of the response given the predictors, which makes testing conditional mean independence more appealing than testing conditional independence. Further, it is very common in practice that some given covariates Z have been known to affect the conditional mean of Y. In this situation, we aim to determine whether X has marginal effect on the conditional mean of Y in the presence of Z, and decide whether X should be included to model the conditional mean of Y along with Z. In general, testing (2) is more useful than testing (1), but requires more careful handling.

We first simplify testing (2) to testing conditional mean independence through a transformation. Let $V = Y - \mathrm{E}(Y|Z) \in \mathbb{R}^q$, and $U = (X, Z) \in \mathbb{R}^{p+r}$. Then $\mathrm{E}(V) = 0$, and $\mathrm{E}(V|U) = \mathrm{E}(Y|X,Z) - \mathrm{E}(Y|Z)$. As a result, we obtain an equivalent hypothesis test to (2) as

$$H_0 : \mathrm{E}(V|U) = \mathrm{E}(V) = 0 \ \ a.s., \quad H_A : o.w. \quad (3)$$

which is conditional mean independence of V given U. Thus, we consider the MDD with U and V to investigate (3). However, there are two problems to solve when we apply MDD to U and V. First, V needs to be estimated since it is unobserved. We will replace V by its estimate $\widehat{V}$ in calculating MDD. Second, we need to confirm that the estimation error of $\widehat{V}$ is negligible, i.e., MDD with $\widehat{V}$ is close enough to that with V, such that $\widehat{V}$ may be used for inference instead of V.

The rest of this paper is organized as follows. In Section 2, we give a brief overview of martingale difference divergence. In Section 3, we estimate V based on the assumption that $\mathrm{E}(Y|Z)$ is a linear function of Z, and prove that the estimation of V does not affect the asymptotic distribution of martingale difference divergence under some regularity conditions. We present simulation results in Section 4, followed by a real data analysis in Section 5[1]. Finally, we summarize our work in Section 6.

The following notation is used throughout this paper. Let $\{(X_i, Y_i, Z_i) : i = 1, \ldots, n\}$ be an i.i.d. sample from the joint distribution $F_{X,Y,Z}$. When A is a matrix, the element of A at row k and column ℓ is denoted by $A(k, \ell)$. When A is a vector, the element of A at index k is denoted by $A(k)$. The Frobenius norm of matrix $A \in \mathbb{R}^{p\times q}$ is denoted by $\|A\|_{\mathrm{F}}$. The Euclidean norm of vector $X \in \mathbb{R}^p$ is denoted by $|X|$. The weighted $\mathcal{L}_2$ norm $\|\cdot\|_w$ of any complex-valued function $\eta(t), t \in \mathbb{R}^p$ is defined by $\|\eta(t)\|_w^2 = \int_{\mathbb{R}^p} |\eta(t)|^2 w(t)\, dt$ where $|\eta(t)|^2 = \eta(t)\overline{\eta(t)}$, $\overline{\eta(t)}$ is the complex conjugate of $\eta(t)$, and $w(t)$ is any positive weight function under which the integral exists. Furthermore, $a.s.$ is an abbreviation of almost surely.

2 MARTINGALE DIFFERENCE DIVERGENCE

Shao and Zhang (2014) proposed martingale difference divergence to capture the conditional mean dependence (in any form) of $Y \in \mathbb{R}^q$ given $X \in \mathbb{R}^p$.

The non-negative martingale difference divergence for X and Y, $\mathrm{MDD}(Y|X)$ is defined by its square

$$\begin{aligned}\mathrm{MDD}^2(Y|X) &= \|\mathrm{E}(Ye^{i\langle s,X\rangle}) - \mathrm{E}(Y)\mathrm{E}(e^{i\langle s,X\rangle})\|_{w_p}^2 \\ &\triangleq \int_{\mathbb{R}^p} |\mathrm{E}(Ye^{i\langle s,X\rangle}) - \mathrm{E}(Y)\mathrm{E}(e^{i\langle s,X\rangle})|^2 w_p(s)\, ds,\end{aligned}$$

where the weight $w_p(s) = c_p|s|^{1+p}$, with $c_p = \frac{\pi^{(1+p)/2}}{\Gamma((1+p)/2)}$, and Γ is the gamma function. If $\mathrm{E}(|X|^2 + |Y|^2) < \infty$, then $\mathrm{MDD}(Y|X) = 0$ if and only if $\mathrm{E}(Y|X) = \mathrm{E}(Y)$ holds $a.s.$

The non-negative empirical martingale difference diver-

[1] See CRAN for an accompanying R package `EDMeasure` (Jin et al., 2018).

gence $\text{MDD}_n(Y|X)$ is analogously defined by

$$\text{MDD}_n^2(Y|X) = \frac{1}{n^2}\sum_{i,j=1}^{n} A_{ij}B_{ij},$$

where $A_{ij} = a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}$, $\bar{a}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}$, $\bar{a}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} a_{ij}$, $\bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij}$, $a_{ij} = |X_i - X_j|$, and similarly for B_{ij} with $b_{ij} = \frac{1}{2}|Y_i - Y_j|^2$.

The consistency and weak convergence of $\text{MDD}_n(Y|X)$ are derived as follows. If $\text{E}(|X| + |Y|^2) < \infty$, we have (i) $\text{MDD}_n(Y|X) \xrightarrow[n\to\infty]{a.s.} \text{MDD}(Y|X)$; (ii) under H_0 : $\text{E}(Y|X) = \text{E}(Y)$ $a.s.$, $n\text{MDD}_n^2(Y|X) \xrightarrow[n\to\infty]{\mathcal{D}} \|\zeta(s)\|_{w_p}^2$, where $\zeta(\cdot)$ is a complex-valued zero-mean Gaussian process whose covariance function depends on $F_{X,Y}$; (iii) under $H_A : o.w.$, $n\text{MDD}_n^2(Y|X) \xrightarrow[n\to\infty]{a.s.} \infty$. Utilizing the nice properties of MDD, we next propose our test for (3).

3 METHODOLOGY

Inspired by the linear assumption to simplify the conditional dependence structure in Fan et al. (2015), we assume that the conditional expectation $\text{E}(Y|Z)$ is a linear function of Z, simplifying the conditional mean dependence structure. As a result, we can decompose Y into the conditional expectation and reminder as

$$Y = \text{E}(Y|Z) + [Y - \text{E}(Y|Z)] \triangleq BZ + V,$$

where $B \in \mathbb{R}^{q\times r}$, $V \in \mathbb{R}^q$. Then we have $\text{E}(V|Z) = 0$, and $\text{E}(V) = 0$. Similarly, the ith sample counterpart is $Y_i = \text{E}(Y_i|Z_i) + V_i \triangleq BZ_i + V_i$, $i = 1, \ldots, n$.

Suppose $\widehat{B}$ is the ordinary least squares (OLS) estimator of B when regressing Y on Z. We will then replace B with $\widehat{B}$ to estimate $\text{E}(Y_i|Z_i)$ as $\widehat{\text{E}}(Y_i|Z_i) = \widehat{B}Z_i$, and V_i as $\widehat{V}_i = Y_i - \widehat{B}Z_i = (B - \widehat{B})Z_i + V_i$. When estimating B via the OLS, Z is implicitly assumed to have full column rank. In case Z is high-dimensional, i.e., $r > n$, we can estimate B by the penalized least squares (PLS) similar to Fan et al. (2015), including ridge (Hoerl and Kennard, 1970) and lasso (Tibshirani, 1996).

We now construct a test for (3) based on $\text{MDD}_n^2(\widehat{V}|U)$ and its counterparts using permutation samples, then calculate the empirical p-value following the permutation in Park et al. (2015). Because the samples are independent, but with an unspecified distribution, permutation tests are a convenient tool for inference. We will later show in Theorem 2 that the asymptotic distribution of $n\text{MDD}_n^2(\widehat{V}|U)$ depends on an unknown underlying distribution, which justifies the use of permutation tests. To measure the conditional mean dependence of V given U, we first compute the test statistic $\text{MDD}_n^2(\widehat{V}|U)$ from the sample $\{(\widehat{V}_i, U_i) : i = 1, \ldots, n\}$, where $U_i = (X_i, Z_i)$. That is, $\text{MDD}_n^2(\widehat{V}|U)$ depends on the i.i.d. sample $\{(X_i, Y_i, Z_i) : i = 1, \ldots, n\}$. Next we draw B permutation samples of size n as $\{(X_i^*, Y_i, Z_i) : i = 1, \cdots, n\}$, where only the sample of X is permuted in order to approximate the sampling distribution. For each permutation sample, we calculate the test statistic $\text{MDD}_{n,b}^2(\widehat{V}|U)$, $b = 1, \cdots, B$. Then the empirical p-value is given by

$$\widehat{p} = \frac{\sum_{b=1}^{B} \mathbf{1}\left\{\text{MDD}_{n,b}^2(\widehat{V}|U) \geq \text{MDD}_n^2(\widehat{V}|U)\right\}}{B}.$$

When H_0 is false, $\text{MDD}_n^2(\widehat{V}|U)$ tends to be large while $\text{MDD}_{n,b}^2(\widehat{V}|U)$ tends to be small. As a result, the empirical p-value is expected to be very small, leading to a rejection of H_0. We name the proposed test linear martingale difference divergence (LinMDD). To justify our LinMDD test, it remains to validate that $\text{MDD}_n^2(\widehat{V}|U)$ is close enough to $\text{MDD}_n^2(V|U)$, i.e., the estimation error in $\widehat{V}$ is negligible for the sampling distribution of the test statistic, focusing on the asymptotic case. To begin with, we introduce some regularity conditions to derive the asymptotic distribution of $\text{MDD}_n^2(\widehat{V}|U)$.

Condition 1. *There exist constants $0 < c_1, c_2, c_3 < \infty$, such that $E(|U_i - U_j|^2) = c_1$, $i \neq j$; $E(|U_i - U_j||U_i - U_k|) = c_2$, $i \neq j \neq k$; $E(|U_i - U_j||U_k - U_\ell|) = c_3$, $i \neq j \neq k \neq \ell$.*

Condition 2. *There exists constant $0 < c_4 < \infty$, such that $E[(Z_i(t) - Z_j(t))^2(Z_i(s) - Z_j(s))^2] \leq c_4$, $i \neq j$, $\forall t, s$.*

Condition 3. *There exists constant $0 < c_5 < \infty$, such that $E[(Z_i(t) - Z_j(t))^2(V_i(s) - V_j(s))^2] \leq c_5$, $i \neq j$, $\forall t, s$.*

Condition 4. $\|\widehat{B} - B\|_F = O_p(n^{-1/2})$.

Remark. Condition 4 can be derived from the bounded density of $|V_i - V_j|$ and non-heavy tails of $Z_i(t)$ and $V_i(t)$ according to Fan et al. (2015) and Fan et al. (2011).

Through a similar derivation to Theorem 2 of Fan et al. (2015), we justify the choice of using $\text{MDD}_n^2(\widehat{V}|U)$ in place of $\text{MDD}_n^2(V|U)$ by the following lemma and theorems. Lemma 1 shows that the difference between $\text{MDD}_n^2(\widehat{V}|U)$ and $\text{MDD}_n^2(V|U)$ is negligible as the sample size increases. The proof of Lemma 1 can be found in Appendix 6.

Lemma 1. *If $Y = BZ + V$ and Conditions 1-4 hold, we have*

$$\mathit{MDD}_n^2(\widehat{V}|U) - \mathit{MDD}_n^2(V|U) = O_p(n^{-3/2}).$$

Consequently, the consistency and weak convergence of $\text{MDD}_n(\widehat{V}|U)$ follow from Lemma 1 and are summarized in Theorem 1 and 2 below.

Theorem 1 (Consistency). *If $Y = BZ + V$ and Conditions 1-4 hold, we have*

$$MDD_n(\widehat{V}|U) \xrightarrow[n\to\infty]{\mathcal{P}} MDD(V|U).$$

Theorem 2 (Weak convergence). *If $Y = BZ + V$ and Conditions 1-4 hold, under H_0, we have*

$$nMDD_n^2(\widehat{V}|U) \xrightarrow[n\to\infty]{\mathcal{D}} \|\zeta(s)\|_{w_p}^2,$$

where $\zeta(\cdot)$ denotes the complex-valued Gaussian random process corresponding to the asymptotic distribution of $nMDD_n^2(V|U)$. Under H_A, we have

$$nMDD_n^2(\widehat{V}|U) \xrightarrow[n\to\infty]{\mathcal{P}} \infty.$$

According to Theorem 1, $\text{MDD}_n(\widehat{V}|U)$ converges to the same population statistic $\text{MDD}(V|U)$ as $\text{MDD}_n(V|U)$, and thus it can serve to measure the conditional mean dependence of V given U. In addition, $n\text{MDD}_n^2(\widehat{V}|U)$ and $n\text{MDD}_n^2(V|U)$ have the same asymptotic distribution stated in Theorem 2, which establishes the effectiveness of LinMDD test, as we approximate the limiting distribution of $n\text{MDD}_n^2(V|U)$ using $n\text{MDD}_n^2(\widehat{V}|U)$ in LinMDD test. In Section 4 and Section 5, we will present the finite-sample performance of our LinMDD test through simulations and a real data example, respectively.

4 SIMULATION STUDIES

To evaluate the performance of our LinMDD test, we adopt the simulation setup in Lavergne and Vuong (2000), and compare our test to the pMDD test (Park et al., 2015), pdCov test (Székely and Rizzo, 2014), and pcov test (Lan et al., 2014) as benchmarks. All tests are implemented as permutation tests with permutation size $B = 500$, in which we only permute the sample of X to approximate the distribution of the test statistic.

We generate data from the underlying model

$$Y = -Z + b \cdot Z^3 + f(X) + \epsilon,$$

where $Z \sim N(0,1)$, $X \sim \mathcal{N}(0,1)$, $\epsilon \sim \mathcal{N}(0,4)$, and Z, X, ϵ are independent. We test the null hypothesis H_0 : $\text{E}(Y|X,Z) = \text{E}(Y|Z)$ $a.s.$ with significance level $\alpha \in \{0.05, 0.1\}$, and examine the empirical size and power of each test. We run 1000 replications with sample size $n \in \{20, 30, 50, 70, 100\}$ for each specific model.

Model 1 (Linear Z, linear X). $b = 0$, $f(X) = cX$ where $c \in \{0, \frac{2}{3}, 1, \frac{3}{2}\}$.

Model 2 (Linear Z, non-linear X). $b = 0$, $f(X) = \sin(c\pi X)$ where $c \in \{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}\}$. We omit $c = 0$ as it is exactly the same as $c = 0$ in Model 1.

From Figure 1, the empirical size of all tests is around 0.05 (0.1). The empirical power of all tests increases as n increases. For the linear X case, the empirical power of all tests is higher when c is larger, since the signal-to-noise ratio increases. Moreover, the empirical power of the LinMDD and pcov tests is consistently higher than that of the other tests, because the linear assumption is valid, and only LinMDD and pcov tests are designed for linear Z. For the non-linear X case, the LinMDD test still outperforms the other tests, while the performance of the pcov test degrades as c increases, because the LinMDD test is designed for non-linear X while pcov test is suitable only for linear X.

Model 3 (Nonlinear Z, linear X). $b = 1$, $f(X) = cX$ where $c \in \{0, \frac{2}{3}, 1, \frac{3}{2}\}$.

Model 4 (Nonlinear Z, non-linear X). $b = 1$, $f(X) = \sin(c\pi X)$ where $c \in \{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}\}$. We omit $c = 0$ as it is exactly the same as $c = 0$ in Model 3.

From Figure 2, the empirical size of all tests is around 0.05 (0.1). For the linear X case, the empirical power of the LinMDD and pcov tests is competitive with but not always higher than that of the other tests. The reason is that the linear dependence of Y on Z is violated while the other tests do not rely it. For the non-linear X case, we similarly find that the performance of the pcov test degrades as c increases. The simulation results show that our LinMDD test achieves competitive and often better performance than the others in these situations. Next, we apply the proposed LinMDD test on a real dataset.

5 FINANCIAL DATA APPLICATION

In finance, the capital asset pricing model (CAPM) was proposed by Sharpe (1964), Lintner (1965), and Mossin (1966) to describe the stock returns through the market risk as

$$r_t = \alpha + \beta_1 m_t,$$

where r_t is the excess stock return (in excess the risk-free return), and m_t is the excess market return at time t. Fama and French (1993) added size and value factors to the CAPM, and proposed the Fama−French three-factor model as

$$r_t = \alpha + \beta_1 m_t + \beta_2 \,\text{SMB}_t + \beta_3 \,\text{HML}_t,$$

where SMB (small minus big) and HML (high minus low) account for stocks with small/big market capitalization and high/low book-to-market ratio, respectively. Fama and French (2015) further added profitability and investment factors to the three-factor model, and extended it to the Fama−French five-factor model as

$$r_t = \alpha + \beta_1 m_t + \beta_2 \,\text{SMB}_t + \beta_3 \,\text{HML}_t + \beta_4 \,\text{RMW}_t + \beta_5 \,\text{CMA}_t,$$

where RMW (robust minus weak) and CMA (conservative minus aggressive) further account for stocks with robust/weak operating profitability and conservative/aggressive investment, respectively.

We collect the annual risk-free returns and Fama−French five factors[2], and the annual returns of Boeing (BA) stock[3] in the past 53 years between 1964 and 2016. The time series and histograms of excessive BA stock returns and Fama−French five factors are depicted in Figure 3.

5.1 CAPM VS. FAMA−FRENCH THREE-FACTOR MODEL

First, we are curious whether the size and value factors should be added to the CAPM, i.e., whether SMB and HML in the Fama−French three-factor model contribute to the expectation of excess stock returns given the market risk. Thus, we test $H_0 : \mathrm{E}(Y|X,Z) = \mathrm{E}(Y|Z)\ a.s.$, where $X_t = (\mathrm{SMB}_t, \mathrm{HML}_t)$, $Y_t = r_t$, and $Z_t = (1, m_t)$.

We apply our LinMDD test to the data with $n = 53$ and $B = 500$. Our p-value is 0.072, while the p-values are 0.012 (pMDD), 0.092 (pdCov) and 0.096 (pcov) using competing tests. As a result, we reject H_0 with significance level $\alpha = 0.1$, and conclude that SMB and HML help determine the excess returns of BA stock in the presence of the market risk. Our results align with the research in finance that the Fama−French three-factor model remarkably outperforms the CAPM in explaining excess stock returns.

5.2 FAMA−FRENCH THREE-FACTOR MODEL VS. FIVE-FACTOR MODEL

Similarly, we are interested in whether the profitability and investment factors should be further added to the Fama−French three-factor model, i.e., whether RMW and CMA in the Fama−French five-factor model contribute to the description of excess stock returns given the other three factors. Hence, we test $H_0 : \mathrm{E}(Y|X,Z) = \mathrm{E}(Y|Z)\ a.s.$, in which $X_t = (\mathrm{RMW}_t, \mathrm{CMA}_t)$, $Y_t = r_t$, and $Z_t = (1, m_t, \mathrm{SMB}_t, \mathrm{HML}_t)$.

We apply our LinMDD test to the data with $n = 53$ and $B = 500$, and its p-value is 0.360, while the p-values are 0.358 (pMDD), 0.878 (pdCov) and 0.768 (pcov) using competing tests. As a result, we fail to reject H_0 with significance level $\alpha = 0.1$, and conclude that RMW and CMA are unable to help determine the excess returns of BA stock in the presence of the other three factors. Our results align with the research in finance that the Fama−French five-factor model has yet to be proven as a significant improvement over the three-factor model in describing excess stock returns.

5.3 FAMA−FRENCH FOUR-FACTOR MODEL VS. FIVE-FACTOR MODEL

Fama and French (2015) showed that the value factor HML becomes redundant when profitability and investment factors are added to the Fama−French three-factor model, because HML is fully captured by its exposures to the other four factors, especially RMW and CMA. To validate this argument, we test $H_0 : \mathrm{E}(Y|X,Z) = \mathrm{E}(Y|Z)\ a.s.$, where $X_t = \mathrm{HML}_t$, $Y_t = r_t$, and $Z_t = (1, m_t, \mathrm{SMB}_t, \mathrm{RMW}_t, \mathrm{CMA}_t)$.

We apply our LinMDD test to the data with $n = 53$ and $B = 500$. Our p-value is 0.218, while the p-values are 0.438 (pMDD), 0.540 (pdCov) and 0.858 (pcov) using competing tests. As a result, we fail to reject H_0 with significance level $\alpha = 0.1$, and conclude that HML cannot help explain the excess returns of BA stock in the presence of the other four factors. Our results demonstrate that HML is redundant for describing excess stock returns in the Fama−French five-factor model.

6 CONCLUSION

In this paper, we propose a new test, LinMDD, for the null hypothesis $H_0 : \mathrm{E}(Y|X,Z) = \mathrm{E}(Y|Z)\ a.s.$ by investigating an equivalent one $H_0 : \mathrm{E}(V|U) = \mathrm{E}(V) = 0\ a.s.$, derived from a transformation involving the conditional expectation. When applying martingale difference divergence (Shao and Zhang, 2014) to test $H_0 : \mathrm{E}(V|U) = \mathrm{E}(V) = 0\ a.s.$, we make two major contributions.

(1) Since V is unobservable, we estimate V based on the assumption that $\mathrm{E}(Y|Z)$ is a linear function of Z, simplifying the conditional mean dependence structure.

(2) We prove that the estimation error in $\widehat{V}$ is negligible for the asymptotic distribution of the test statistic. Thus, we can replace V with $\widehat{V}$ in the test statistic for inference in large samples.

We implement the LinMDD test as a permutation test following Park et al. (2015), and compare it with existing tests in various simulation studies. The LinMDD test consistently outperforms existing tests when its linear assumption is valid, and it achieves competitive results with existing tests even when its linear assumption is violated.

[2]Download data at http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

[3]Download data using `get.hist.quote` in the R package `tseries` (Trapletti and Hornik, 2017).

To illustrate the practical value of the LinMDD test, we compare the CAPM, the Fama−French three-factor and five-factor models by applying LinMDD test to the financial data. We find that the Fama−French three-factor outperforms the CAPM, while the Fama−French five-factor is not a significant improvement over the three-factor model when explaining the excess annual returns of a major stock. Moreover, we validate the statement that the value factor is redundant in the Fama−French five-factor model (Fama and French, 2015) using the LinMDD test.

The relaxation of the linear assumption is an important topic for future research. Our method will become more general if the linear assumption of conditional mean dependence can be generalized to a non-linear one, using non-parametric regression (local regression, splines) instead of linear regression in the estimation of conditional mean. In addition, the high-dimensional setting regarding Z where $r > n$ is an interesting direction to consider as well.

APPENDIX

PROOF OF LEMMA 1

Proof. We define T

$$
\begin{aligned}
&= n\text{MDD}_n^2(\widehat{V}|U) - n\text{MDD}_n^2(V|U) \\
&= \frac{1}{2n}\sum_{i,j}[(F_{ij} - \frac{1}{n}\sum_k F_{kj} - \frac{1}{n}\sum_k F_{ik} + \frac{1}{n^2}\sum_{k,\ell} F_{k\ell}) \\
&\times (E_{ij} - \frac{1}{n}\sum_k E_{kj} - \frac{1}{n}\sum_k E_{ik} + \frac{1}{n^2}\sum_{k,\ell} E_{k\ell})],
\end{aligned}
$$

where $F_{ij} = |\widehat{V}_i - \widehat{V}_j|^2 - |V_i - V_j|^2$, $E_{ij} = |U_i - U_j|$.

We apply Taylor expansion to $|\widehat{V}_t - \widehat{V}_s|^2$ at $V_t - V_s$ in terms of $f(x) = x^T x$, $f'(x) = 2x^T$, then there exists $\lambda \in (0, 1)$, such that F_{ij}

$$
\begin{aligned}
&= 2[\lambda(\widehat{V}_i - \widehat{V}_j) + (1-\lambda)(V_i - V_j)]^T(\widehat{V}_i - \widehat{V}_j - V_i + V_j) \\
&= 2[\lambda(Z_i - Z_j)^T(B - \widehat{B})^T(B - \widehat{B})(Z_i - Z_j) \\
&+ (V_i - V_j)^T(B - \widehat{B})(Z_i - Z_j)].
\end{aligned}
$$

Thus, we have $T = T_1 + T_2$, where T_1

$$
\begin{aligned}
&= \frac{\lambda}{n}\sum_{i,j}[(G_{ij} - \frac{1}{n}\sum_k G_{kj} - \frac{1}{n}\sum_k G_{ik} + \frac{1}{n^2}\sum_{k,\ell} G_{k\ell}) \\
&\times (E_{ij} - \frac{1}{n}\sum_k E_{kj} - \frac{1}{n}\sum_k E_{ik} + \frac{1}{n^2}\sum_{k,\ell} E_{k\ell})],
\end{aligned}
$$

$$
G_{ij} = (Z_i - Z_j)^T(B - \widehat{B})^T(B - \widehat{B})(Z_i - Z_j),
$$

and T_2

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i,j}[(H_{ij} - \frac{1}{n}\sum_k H_{kj} - \frac{1}{n}\sum_k H_{ik}\frac{1}{n^2}\sum_{k,\ell} H_{k\ell}) \\
&\times (E_{ij} - \frac{1}{n}\sum_k E_{kj} - \frac{1}{n}\sum_k E_{ik} + \frac{1}{n^2}\sum_{k,\ell} E_{k\ell})],
\end{aligned}
$$

$$
H_{ij} = (V_i - V_j)^T(B - \widehat{B})(Z_i - Z_j).
$$

First, we will show (i) $T_1 = O_p(n^{-1})$.

After a simple calculation, we have

$$
\begin{aligned}
&\frac{1}{n}\sum_{i,j}(G_{ij} - \frac{1}{n}\sum_k G_{kj} - \frac{1}{n}\sum_k G_{ik} + \frac{1}{n^2}\sum_{k,\ell} G_{k\ell})E_{ij} \\
&= tr[\frac{1}{n}\sum_{i,j}|U_i - U_j|(G_{ij} - \frac{1}{n}\sum_k G_{kj} - \frac{1}{n}\sum_k G_{ik} \\
&+ \frac{1}{n^2}\sum_{k,\ell} G_{k\ell})] \\
&= tr[(B - \widehat{B})^T(B - \widehat{B})M],
\end{aligned}
$$

where $M = \frac{1}{n}\sum_{i,j}|U_i - U_j|S_{ij}$, and

$$
S_{ij} = R_{ij} - \frac{1}{n}\sum_k R_{kj} - \frac{1}{n}\sum_k R_{ik} + \frac{1}{n^2}\sum_{k,\ell} R_{k\ell},
$$

$R_{ij} = (Z_i - Z_j)(Z_i - Z_j)^T$, $R_{ij} = R_{ji}$, $S_{ij} = S_{ji}$, then

$$
\begin{aligned}
&\text{E}[(M(t,s))^2] \\
&= \text{E}[\frac{1}{n^2}(\sum_{i,j}|U_i - U_j|S_{ij}(t,s))^2] \\
&= \text{E}\{\text{E}[\frac{1}{n^2}(\sum_{i,j}|U_i - U_j|S_{ij}(t,s))^2|U_i, \forall i]\} \\
&= \text{E}[\frac{2c_1}{n^2}\sum_{i\neq j}(S_{ij}(t,s))^2 \\
&+ \frac{2c_2}{n^2}\sum_{i\neq j\neq k}(S_{ij}(t,s)S_{ik}(t,s) + S_{ij}(t,s)S_{kj}(t,s)) \\
&+ \frac{c_3}{n^2}\sum_{i\neq j\neq k\neq \ell} S_{ij}(t,s)S_{k\ell}(t,s)],
\end{aligned}
$$

where $c_1 = \text{E}(|U_i - U_j|^2)$, $i \neq j$; $c_2 = \text{E}(|U_i - U_j||U_i - U_k|)$, $i \neq j \neq k$; $c_3 = \text{E}(|U_i - U_j||U_k - U_\ell|)$, $i \neq j \neq k \neq \ell$.

Considering that $\text{E}[(R_{ij}(t,s))^2] = \text{E}[(Z_i - Z_j)_t^2(Z_i - Z_j)_s^2] \leq c_4$, $i \neq j$, $\forall t, s$, we have $\text{E}[(R_{ij}(t,s))^2] = O(1)$, which implies $\text{E}[(S_{ij}(t,s))^2] = O(1)$, and thus $\text{E}[\frac{1}{n^2}\sum_{i\neq j}(S_{ij}(t,s))^2] = O(1)$.

After a simple calculation, we have $\sum_i S_{ij}(t,s) = 0$, $\sum_j S_{ij}(t,s) = 0$, $\sum_i \sum_j S_{ij}(t,s) = 0$, and

$$\begin{aligned}
&\sum_{i\neq j\neq k} S_{ij}(t,s)S_{ik}(t,s) \\
&= \sum_i (S_{ii}(t,s))^2 - \sum_{i\neq j}(S_{ij}(t,s))^2, \\
&\sum_{i\neq j\neq k} S_{ii}(t,s)S_{jk}(t,s) \\
&= \sum_i (S_{ii}(t,s))^2 - \sum_{i\neq j} S_{ii}(t,s)S_{jj}(t,s), \\
&\sum_{i\neq j\neq k\neq \ell} S_{ij}(t,s)S_{k\ell}(t,s) \\
&= -2\sum_{i\neq j\neq k}[S_{ii}(t,s)S_{jk}(t,s) + S_{ij}(t,s)S_{ik}(t,s) \\
&+ S_{ij}(t,s)S_{kj}(t,s)] \\
&- \sum_{i\neq j}[4S_{ii}(t,s)S_{ij}(t,s) + S_{ii}(t,s)S_{jj}(t,s) \\
&+ 2(S_{ij}(t,s))^2] - \sum_i (S_{ii}(t,s))^2,
\end{aligned}$$

we have

$$\begin{aligned}
\mathrm{E}[\frac{1}{n^2}\sum_{i\neq j\neq k} S_{ij}(t,s)S_{ik}(t,s)] = O(1), \\
\mathrm{E}[\frac{1}{n^2}\sum_{i\neq j\neq k} S_{ij}(t,s)S_{kj}(t,s)] = O(1), \\
\mathrm{E}[\frac{1}{n^2}\sum_{i\neq j\neq k} S_{ii}(t,s)S_{jk}(t,s)] = O(1), \\
\mathrm{E}[\frac{1}{n^2}\sum_{i\neq j\neq k\neq \ell} S_{ij}(t,s)S_{k\ell}(t,s)] = O(1).
\end{aligned}$$

Therefore, $\mathrm{E}[(M(t,s))^2] = O(1)$.

Applying Chebyshev's inequality to $M(t,s)$, we have

$$P(|M(t,s) - \mu| \geq k\sigma) \leq 1/k^2,$$

where $\mu = \mathrm{E}[M(t,s)]$, $\sigma^2 = \mathrm{Var}[M(t,s)]$. As a result, $M(t,s) = O_p(1)$.

Given that $\|\widehat{B} - B\|_F = O_p(n^{-1/2})$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i,j}(G_{ij} - \frac{1}{n}\sum_k G_{kj} - \frac{1}{n}\sum_k G_{ik} + \frac{1}{n^2}\sum_{k,\ell} G_{k\ell})E_{ij} \\
&= tr[(B - \widehat{B})^T(B - \widehat{B})M] \\
&= pq^2 O_p(n^{-1})O_p(1) \\
&= O_p(n^{-1}).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i,j}(G_{ij} - \frac{1}{n}\sum_k G_{kj} - \frac{1}{n}\sum_k G_{ik} + \frac{1}{n^2}\sum_{k,\ell} G_{k\ell})E_{kj}, \\
&\frac{1}{n}\sum_{i,j}(G_{ij} - \frac{1}{n}\sum_k G_{kj} - \frac{1}{n}\sum_k G_{ik} + \frac{1}{n^2}\sum_{k,\ell} G_{k\ell})E_{ik}, \\
&\frac{1}{n}\sum_{i,j}(G_{ij} - \frac{1}{n}\sum_k G_{kj} - \frac{1}{n}\sum_k G_{ik} + \frac{1}{n^2}\sum_{k,\ell} G_{k\ell})E_{k\ell}
\end{aligned}$$

are all $O_p(n^{-1})$. Therefore, $T_1 = O_p(n^{-1})$.

Analogous to (i), we can show (ii) $T_2 = O_p(n^{-1/2})$. The only differences are

$$\begin{aligned}
&\frac{1}{n}\sum_{i,j}(H_{ij} - \frac{1}{n}\sum_k H_{kj} - \frac{1}{n}\sum_k H_{ik} + \frac{1}{n^2}\sum_{k,\ell} H_{k\ell})E_{ij} \\
&= tr[(B - \widehat{B})M],
\end{aligned}$$

where M is defined similarly with $R_{ij} = (Z_i - Z_j)(V_i - V_j)^T$, and $\mathrm{E}[(R_{ij}(t,s))^2] = \mathrm{E}[(Z_i - Z_j)_t^2(V_i - V_j)_s^2] \leq c_5$, $i \neq j$, $\forall t, s$, and

$$\begin{aligned}
&\frac{1}{n}\sum_{i,j}(H_{ij} - \frac{1}{n}\sum_k H_{kj} - \frac{1}{n}\sum_k H_{ik} + \frac{1}{n^2}\sum_{k,\ell} H_{k\ell})E_{ij} \\
&= tr[(B - \widehat{B})M] \\
&= pqO_p(n^{-1/2})O_p(1) \\
&= O_p(n^{-1/2}),
\end{aligned}$$

and therefore $T_2 = O_p(n^{-1/2})$.

As a conclusion, $T = T_1 + T_2 = O_p(n^{-1/2})$. □

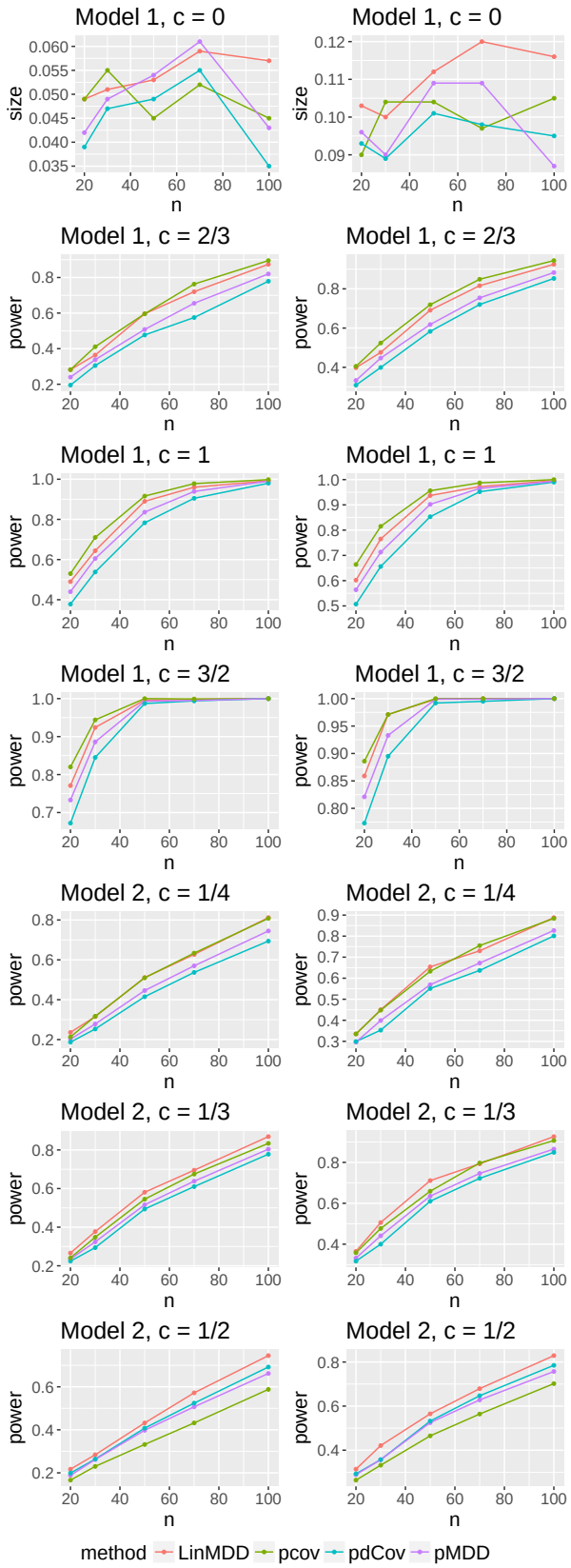

Figure 1: Empirical size and power of 1000 replications with $B = 500$ for Model 1 & 2.

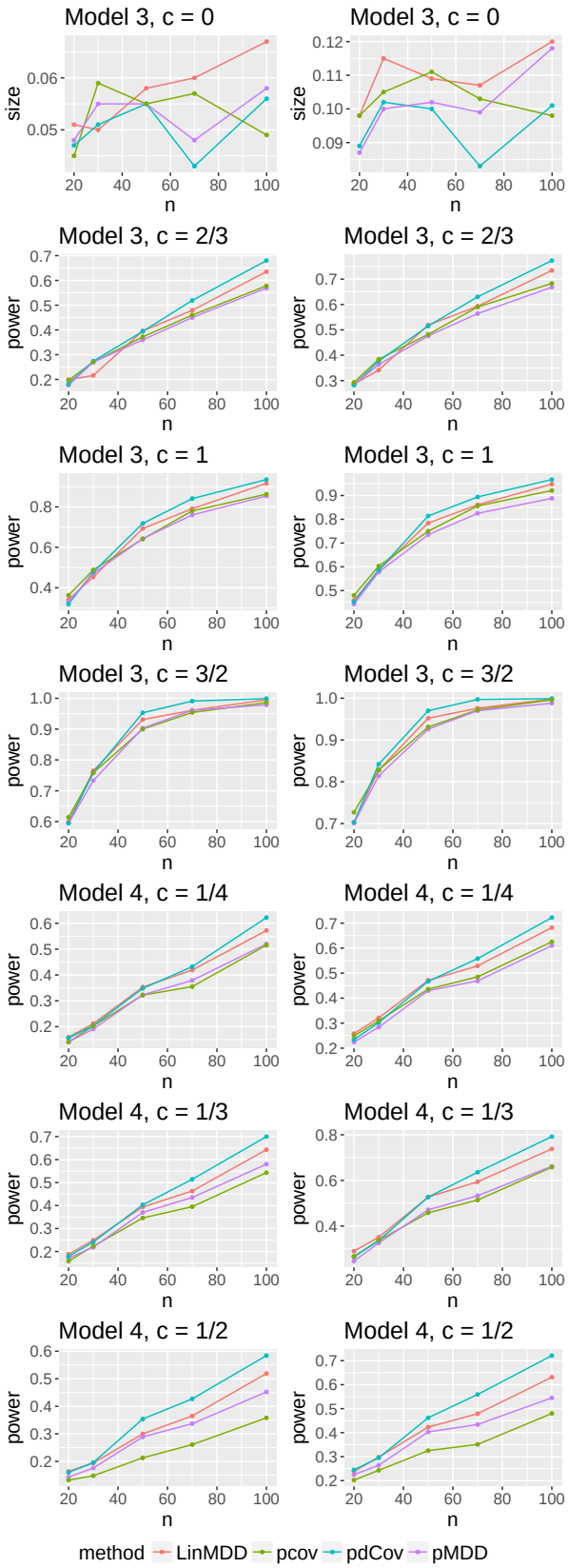

Figure 2: Empirical size and power of 1000 replications with $B = 500$ for Model 3 & 4.

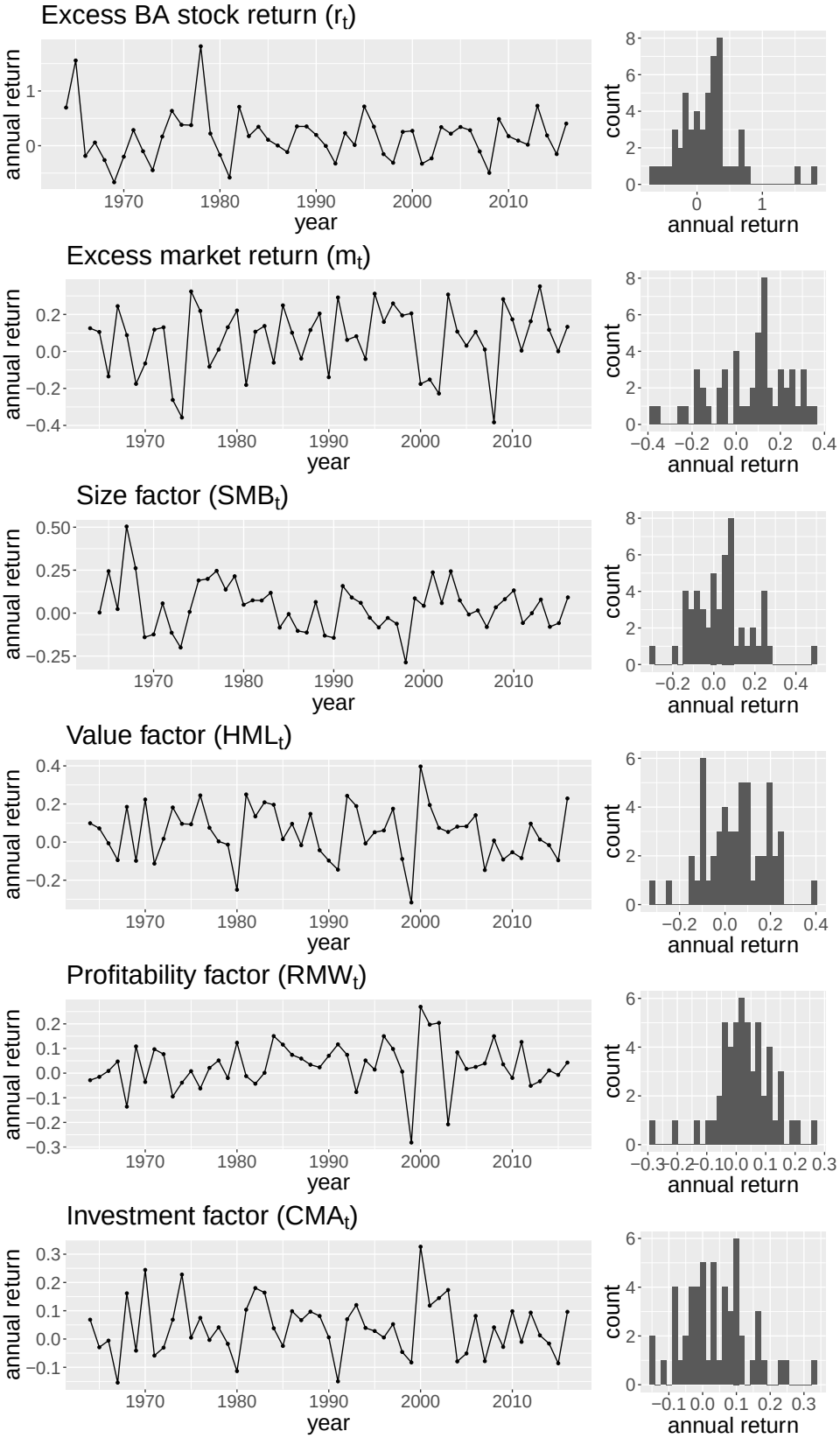

Figure 3: Time series and histograms of excess BA stock returns (r_t), excess market returns (m_t), size factors (SMB_t), value factors (HML_t), profitability factors (RMW_t), and investment factors (CMA_t) between 1964 and 2016.

References

Y. Aït-Sahalia, P. J. Bickel, and T. M. Stoker. Goodness-of-fit tests for kernel regression with an application to option implied volatilities. *Journal of Econometrics*, 105(2):363–412, 2001.

S. Chatterjee and A. S. Hadi. *Regression analysis by example*. John Wiley & Sons, 2015.

B. Chen and Y. Hong. Testing for the markov property in time series. *Econometric Theory*, 28(1):130–178, 2012.

R. D. Cook and B. Li. Dimension reduction for conditional mean in regression. *Annals of Statistics*, pages 455–474, 2002.

E. F. Fama and K. R. French. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56, 1993.

E. F. Fama and K. R. French. A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22, 2015.

J. Fan, Y. Liao, and M. Mincheva. High dimensional covariance matrix estimation in approximate factor models. *Annals of statistics*, 39(6):3320, 2011.

J. Fan, Y. Feng, and L. Xia. A conditional dependence measure with applications to undirected graphical models. *arXiv preprint arXiv:1501.01617*, 2015.

Y. Fan and Q. Li. Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica: Journal of the econometric society*, pages 865–890, 1996.

L. Gan, N. N. Narisetty, and F. Liang. Bayesian regularization for graphical models with unequal shrinkage. *Journal of the American Statistical Association*, (just-accepted), 2018.

C. Hiemstra and J. D. Jones. Testing for linear and nonlinear granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5):1639–1664, 1994.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Y. Hong, Y. Liu, and S. Wang. Granger causality in risk and detection of extreme risk spillover between financial markets. *Journal of Econometrics*, 150(2):271–287, 2009.

Z. Jin and D. S. Matteson. Generalizing distance covariance to measure and test multivariate mutual dependence. *arXiv preprint arXiv:1709.02532*, 2017.

Z. Jin and D. S. Matteson. Independent component analysis via energy-based and kernel-based mutual dependence measures. *arXiv preprint arXiv:1805.06639*, 2018.

Z. Jin, S. Yao, D. S. Matteson, and X. Shao. *EDMeasure: Energy-Based Dependence Measures*, 2018. R package version 1.2.

W. Lan, H. Wang, and C.-L. Tsai. Testing covariates in high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, 66(2):279–301, 2014.

P. Lavergne and Q. Vuong. Nonparametric significance testing. *Econometric Theory*, 16(4):576–601, 2000.

R. Li, W. Zhong, and L. Zhu. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139, 2012.

Z. Li and T. H. McCormick. An expectation conditional maximization approach for gaussian graphical models. *arXiv preprint arXiv:1709.06970*, 2017.

Z. R. Li, T. H. McCormick, and S. J. Clark. Bayesian joint spike-and-slab graphical lasso. *arXiv preprint arXiv:1805.07051*, 2018.

J. Lintner. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The review of economics and statistics*, pages 13–37, 1965.

J. Mossin. Equilibrium in a capital asset market. *Econometrica: Journal of the econometric society*, pages 768–783, 1966.

T. Park, X. Shao, and S. Yao. Partial martingale difference correlation. *Electronic Journal of Statistics*, 9(1): 1492–1517, 2015.

H. Raïssi et al. Testing linear causality in mean when the number of estimated parameters is high. *Electronic journal of statistics*, 5:507–533, 2011.

X. Shao and J. Zhang. Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, 109(507):1302–1318, 2014.

W. F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442, 1964.

L. Su and H. White. A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*, 141(2):807–834, 2007.

G. J. Székely and M. L. Rizzo. Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, 42(6):2382–2412, 2014.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.

Y. Tang, H. J. Wang, and E. Barut. Testing for the presence of significant covariates through conditional marginal regression. *Biometrika*, 2017.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

A. Trapletti and K. Hornik. *tseries: Time Series Analysis and Computational Finance*, 2017. R package version 0.10-42.

X. Wang, W. Pan, W. Hu, Y. Tian, and H. Zhang. Conditional distance correlation. *Journal of the American Statistical Association*, 110(512):1726–1734, 2015.

X. Yan and J. Bien. Rare feature selection in high dimensions. *arXiv preprint arXiv:1803.06675*, 2018.

X. Yan, J. Bien, et al. Hierarchical sparse modeling: A choice of two group lasso formulations. *Statistical Science*, 32(4):531–560, 2017.

B. Zhang, N. Mohammed, V. Dave, and M. A. Hasan. Feature selection for classification under anonymity constraint. *arXiv preprint arXiv:1512.07158*, 2015.